



## 저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

의학박사 학위논문

Development of Survival Prediction Model for the  
Korean Disease-free Lung Cancer Survivors using  
Patient Reported Outcome variables: application to  
Cox proportional hazard regression model and diverse  
machine learning algorithms

환자 보고 성과 지표를 활용한 한국인 폐암 무병  
생존자 생존 예측 모형 개발

— Cox 비례 위험 모델 및 다양한 머신 러닝 알고리즘을 적용하여—

2018 년 02 월

서울대학교 대학원

의과학과 (시스템정보맞춤의학) 전공

심 진 아



Thesis of the Degree of Doctor of Philosophy  
in Medical Sciences

환자 보고 성과 지표를 활용한 한국인 폐암 무병  
생존자 생존 예측 모형 개발

-Cox 비례 위험 모델 및 다양한 머신 러닝 알고리즘을 적용하여-

Development of Survival Prediction Model for the Korean  
Disease-free Lung Cancer Survivors using Patient Reported  
Outcome variables: application to Cox proportional hazard  
regression model and diverse machine learning algorithms

**February 2018**

**Department of Bio-medical Sciences,  
Seoul National University, College of Medicine  
Sim Jin-ah**



Development of Survival Prediction Model for the Korean  
Disease-free Lung Cancer Survivors using Patient Reported  
Outcome variables: application to Cox proportional hazard  
regression model and diverse machine learning algorithms

by

**Jin-ah Sim**

A thesis submitted to the Department of Bio–medical  
Sciences in partial fulfillment of the requirements for the  
Degree of Doctor of Philosophy in Medical Science at  
Seoul National University College of Medicine

February 2018

Approved by Thesis Committee:

Professor \_\_\_\_\_ Chairman

Professor \_\_\_\_\_ Vice chairman

Professor \_\_\_\_\_

Professor \_\_\_\_\_

Professor \_\_\_\_\_

환자 보고 성과 지표를 활용한 한국인 폐암 무병  
생존자 생존 예측 모형 개발

-Cox 비례 위험 모델 및 다양한 머신 러닝 알고리즘을 적용하여-

지도교수 윤 영 호

이 논문을 의학박사 학위논문으로 제출함

2018년 2월

서울대학교 대학원  
의과학과 (시스템정보맞춤의학)  
심 진 아

심진아의 의학박사 학위논문을 인준함

2018 년 02월

위 원 장	김 주 한	(인)
부위원장	윤 영 호	(인)
위 원	한 서 경	(인)
위 원	박 상 민	(인)
위 원	김 영 애	(인)

# ABSTRACT

**Introduction:** The prediction of lung cancer survival is a crucial factor for successful cancer survivorship and follow-up planning. The principal objective of this study is to construct a novel Korean prognostic model of 5-year survival within lung cancer disease-free survivors using socio-clinical and HRQOL variables and to compare its predictive performance with the prediction model based on the traditional known clinical variables. Diverse techniques such as Cox proportional hazard model and machine learning technologies (MLT) were applied to the modeling process.

**Methods:** Data of 809 survivors, who underwent lung cancer surgery between 1994 and 2002 at two Korean tertiary teaching hospitals, were used. The following variables were selected as independent variables for the prognostic model by using literature reviews and univariate analysis: clinical and socio-demographic variables, including age, sex, stage, metastatic lymph node and income; health related quality of life (HRQOL) factors from the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30; Quality of Life Questionnaire Lung Cancer Module; Hospital Anxiety and Depression Scale, and Post-traumatic Growth Inventory. Survivors' body mass index before a surgery and physical activity were also chosen. The three prediction modeling features sets included 1) only clinical and socio-demographic variables, 2) only HRQOL and lifestyle factors, and 3) variables from feature set 1 and 2 considered altogether. For each feature set, three Cox proportional hazard regression model were constructed and compared among each other by evaluating their performance in terms of discrimination and calibration ability using the C-statistic and Hosmer-



Lemeshow chi-square statistics. Further, four machine learning algorithms using decision tree (DT), random forest (RF), bagging, and adaptive boosting (AdaBoost) were applied to three feature sets and compared with the performances of one another. The performance of the derived predictive models based on MLTs were internally validated by K-fold cross-validation.

**Results:** In the Cox modeling, Model Cox-3 (based on Feature set 3: HRQOL factors added into clinical and socio-demographic variables) showed the highest area under curve (AUC = 0.809) compared with two other Cox regression (Cox-1, 2). When we applied the modeling methods into all other MLT based models, the most effective models were Model DT-3 from DT, Model RF-3 from RF, Model Bag-3 from Bagging, Model AdaBoost-3 from AdaBoost techniques, showing the highest accuracy for each of MLT. Model RF-3, Model Bag-3, Model AdaBoost-3 showed the highest accuracy even after k-fold cross-validation were conducted.

**Conclusions:** Considering that the HRQOLs were added with clinical and socio-demographic variables, the proposed model proved to be useful based on the Cox model or we can apply MLT algorithms in the prediction of lung cancer survival. Improved accuracy for lung cancer survival prediction model has the potential to help clinicians and survivors make more meaningful decisions about future plans and their support to cancer care.

-----

**Keywords:** Cancer, Survival, Prediction, HRQOL, Machine Learning

**Student number:** 2015-30606

# CONTENTS

Abstract .....	i
Contents .....	iii
List of Figures.....	vii
LIST of Tables .....	ix
<b>I. INTRODUCTION .....</b>	<b>1</b>
A. Background .....	1
1. Lung cancer statistics.....	1
2. The importance of suggesting survival prediction model to cancer survivors .....	4
3. HRQOL and lifestyle measurement as important predictors for lung cancer survival .....	5
4. Traditional survival analysis versus machine learning techniques (MLTs).....	7
B. Hypothesis and objectives .....	10
1. Hypothesis.....	10
2. Objectives .....	10
<b>II. MATERIALS AND METHODS .....</b>	<b>12</b>
A. Study subjects .....	12
1. Subject selection.....	12
2. Data collection .....	13
2.1. Socio-demographic and clinical variables .....	16
2.2. Patient lifestyle characteristics.....	17
3. Study process .....	21

B. Prognostic variables' selection and data preprocessing.....	22
1. Prognostic variables' selection .....	22
1.1. Literature review for the selection of candidate predictors .....	22
1.2. Grading the evidence and mapping into the conceptual framework .....	25
1.3. Examination of prognosis variables' selection from statistical analyses .....	28
2. Data preprocessing .....	29
2.1. Data cleaning, missing imputation .....	29
2.2. Test of multi-collinearity .....	29
2.3. Decisions of cut-off points.....	30
2.4. Data sampling for data balancing, SMOTE.....	31
2.5. Data splitting (holdout strategy) .....	32
C. Model development.....	33
1. Cox model development .....	34
3. Random forest model.....	38
4. Bagging (bootstrap aggregating) .....	40
5. Adaptive boosting (AdaBoost).....	42
D. Model validation .....	44
1. Model validation for Cox model .....	44
1.1. Discrimination for Cox model .....	44
1.2. Calibration for Cox model.....	44
2. Model validation of other MLTs.....	45
3. K-fold Cross Validation for MLT based prediction models to avoid over- fitting .....	46
<b>III. RESULTS.....</b>	<b>49</b>

A. Literature review for selection of candidate predictors .....	49
1. Selection of candidate prognostic factors with literature review.....	49
2. Model constructing feature sets with selecting prognostic factors .....	52
B. Baseline characteristics .....	53
1. Demographics of participants' characteristics and survival data.....	53
2. Candidate selection from statistical analyses .....	55
2.1. Univariate analysis of HRQOL mean scores between non-event and event groups .....	55
2.2. Univariate analysis of BMI, weight change, and MET of lung cancer survivors.....	59
3. Final candidate variable selection for phased modeling .....	61
4. Result of data preprocessing .....	63
4.1. Missing imputation.....	63
C. Model development.....	65
1. Cox model development .....	66
1.1. Prediction model based on Cox regression analysis.....	68
1.2. Final prediction model equation for Cox models .....	72
2. Decision tree model development .....	73
2.1. Assessment of the relative importance and model developing .....	73
2.2. Selecting CP value for decision tree pruning using “rpart” packages .....	75
3. Random forest model development .....	77
4. Bagged decision tree model development.....	79
5. AdaBoost model development .....	80
6. Developed models applied with MLTs .....	82

D. Model validation and performance .....	89
1. Cox proportional hazard ratio model internal validation .....	89
1.1. Discrimination .....	89
1.2. Calibration .....	92
2. Comparison model performance of Cox model and other MLTs .....	96
<b>IV. DISCUSSION .....</b>	<b>106</b>
A. Literature review for selection of candidate predictors .....	107
B. Model development using Cox and other MLTs .....	109
C. Model validation of Cox regression model and application of the predictive models to other MLT based models .....	112
D. Clinical and practical implications .....	114
E. Strengths and limitations of this study .....	117
<b>CONCLUSION .....</b>	<b>119</b>
<b>REFERENCES .....</b>	<b>120</b>
국문 초록 .....	134
<b>APPENDIX .....</b>	<b>136</b>

# LIST OF FIGURES

Figure I -1. The 10 leading types of cancers' estimated new cases and deaths by both sexes in 2017 .....	2
Figure I -2. Trends in age standardized rates in lung cancer incidence and mortality in Korea based on the joint point regression.....	3
Figure I -3. Classification of survival analysis methods.....	9
Figure I -4. Hypothetical series of tasks diagram of the current study .....	11
Figure II-1. Selection of eligible study subjects in the current study .....	14
Figure II-2. Overall survival (OS) data structure.....	20
Figure II-3. Prediction model development and validation process.....	21
Figure II-4. Stepwise scoring and grading procedure applied to assess the quality of evidence.....	27
Figure II-5. Comparison of balancing methods .....	31
Figure II-6. Hold-out sampling method to avoid over-fitting problems.....	32
Figure II-7. Random forest algorithms .....	39
Figure II-8. Bagging data splitting procedure.....	41
Figure II-9. AdaBoost model algorithms.....	42
Figure II-10. Confusion matrices for the training dataset and test samples.....	45
Figure II-11. <i>k</i> -fold cross validation ( <i>k</i> =5). .....	48
Figure III-1. Flow of selection of candidate prognostic factors from systematic review .....	50
Figure III-2. Schematic diagram of candidate prognostic factors mapped with the bio-psychological framework based on the International Classification of Functioning, Disability and Health (ICF).....	51
Figure III-3. Plots of decision tree models .....	76
Figure III-4. Random forest variable importance plots .....	78
Figure III-5. Out-of-bag (OOB) error rate according to number of bootstraps.....	79
Figure III-6. ROC plot of Cox regression model in development set .....	90
Figure III-7. ROC plot of Cox regression model in validation set.....	91
Figure III-8. Calibration plot of lung cancer prediction model Cox-1 .....	93
Figure III-9. Calibration plot of lung cancer survival prediction model Cox-2 .....	93

Figure III-10. Calibration of lung cancer prediction model Cox-3 .....	94
Figure III-11. AUC curve comparison of lung cancer prediction models based on decision tree .....	97
Figure III-12. AUC curve comparison of lung cancer prediction models based on random forest model.....	98
Figure III-13. AUC curve comparison of lung cancer prediction models based on Bagging techniques .....	99
Figure III-14. AUC curve comparison of lung cancer prediction models based on Bagging techniques .....	100
Figure III-15. Rader chart of performance of three models based on Cox model .....	103
Figure III-16. Rader chart of performance of three models based on DT model.....	103
Figure III-17. Rader chart of performance of three models based on RF model .....	104
Figure III-18. Rader chart of performance of three models based on bagging model.....	104
Figure III-19. Rader chart of performance of three models based on AdaBoost model.....	105

# LIST OF TABLES

Table II-1. Past medical history information and composition of questionnaire .....	15
Table II-2. Patient intervention comparison outcome strategy questions .....	24
Table III-1. Comparison of clinico-pathologic and socio-demographic characteristics between the event and no-event groups .....	54
Table III-2. Comparison of EORTC QLQ-C30 HRQOL factors between the event and no-event groups .....	56
Table III-3. Comparison of EORTC QLQ-LC13 HRQOL factors between the event and no-event groups .....	57
Table III-4. Comparison of PTGI and HADS factors between the event and no-event groups .....	58
Table III-5. Comparison of lifestyle factors between the event and no-event groups .....	60
Table III-6. Final candidate variables from both literature review and statistical analyses .....	62
Table III-7. Comparison of original data structure of alive and death groups with SMOTE data .....	64
Table III-8. Possible models in phased cox modeling for lung cancer survivors .....	67
Table III-9. Lung cancer survivors' mortality prediction model Cox-1 .....	69
Table III-10. Lung cancer survival prediction model Cox-2 .....	70
Table III-11. Lung cancer survival prediction model Cox-3 .....	71
Table III-12. Importance of prognostic factors by normalized mutual information index of DT models .....	74
Table III-13. Importance of prognostic factors by normalized mutual information index .....	81
Table III-14. Selected important variables based on Cox models .....	84
Table III-15. Selected important variables based on DT models .....	85
Table III-16. Selected important variables based on random forest models .....	86
Table III-17. Selected important variables based on Bagging models .....	87
Table III-18. Selected important variables based on AdaBoost models .....	88
Table III-19. C-statistic and Hosmer-Lemeshow type chi-square test for lung cancer	



survival prediction models for development and validation sets .....	95
Table III-20. Performance comparison of three data mining algorithms based on the Cox models .....	101
Table III-21. Performance comparison of three data mining algorithms based on the DT models .....	101
Table III-22. Performance comparison of three data mining algorithms based on the RF models.....	101
Table III-23. Performance comparison of three data mining algorithms based on the Bagging models .....	102
Table III-24. Performance comparison of three data mining algorithms based on the AdaBoost models .....	102

# I. INTRODUCTION

## *A. Background*

### **1. Lung cancer statistics**

Cancer is a major health problem in Republic of Korea, as it is the most important cause of death since 1980s [1]. Annually, one in four individuals die among 200,000 patients newly diagnosed with cancer in Korea [2, 3]. Among all types of cancers, lung cancer has been the most common cancer for several decades globally [4, 5], estimated at 26,093 cases (11.8% of the total) and accounted for the highest proportion of estimated cancer mortality with nearly one in four deaths (18,713 cases, 23.3% of the total) in 2017 [6, 7]. (Figure I -1) In Republic of Korea, lung cancer is the most common cause of cancer death (crude rate (CR) [7, 8], 52.5/100,000; age-standardized rate (ASR), 34.0/100,000 for men and CR, 18.3/100,000; ASR, 8.1/100,000 for women in 2014) estimated to account for 22.05% of all cancer deaths [9-11].

Fortunately, due to advances in early detection and improved treatment strategies, cancer mortality has decreased [4, 5, 12, 13]. (Appendix Figure 1). Though the increase in ASR for lung cancer with an annual percent change (APC) did not show significant changes, it was determined that Korean age-adjusted cancer mortality for lung cancer has been decreasing by 3.4% annually since 2012 (Figure I -2-(A)). The rates have been decreasing at the rate of 2.5% annually among males since 2012 (Figure I -2-(B)). Considering that only one straight line

is displayed among females (Figure I -2-(C)), one can interpret the 2000–2016 trend for female age-adjusted lung cancer mortality as being constant (in this case, consistently decreasing) throughout the 16-year period studied. This progress implies that a growing number of patients may be gradually freed of cancer and considering this, cancer may be managed as a chronic illness requiring long-term surveillance [14, 15].

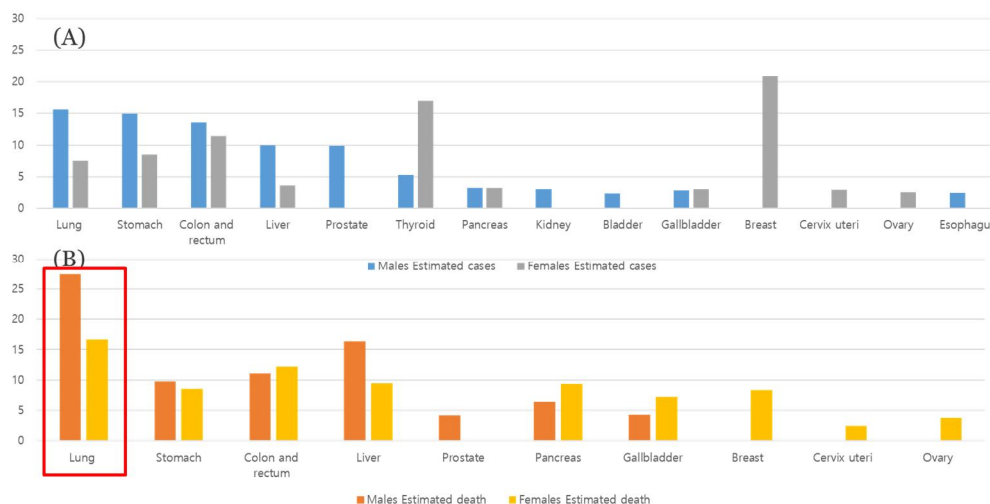


Figure I -1. The 10 leading types of cancers' estimated new cases and deaths by both sexes in 2017  
(A) Estimated new cases. (B) Estimated deaths.[7]

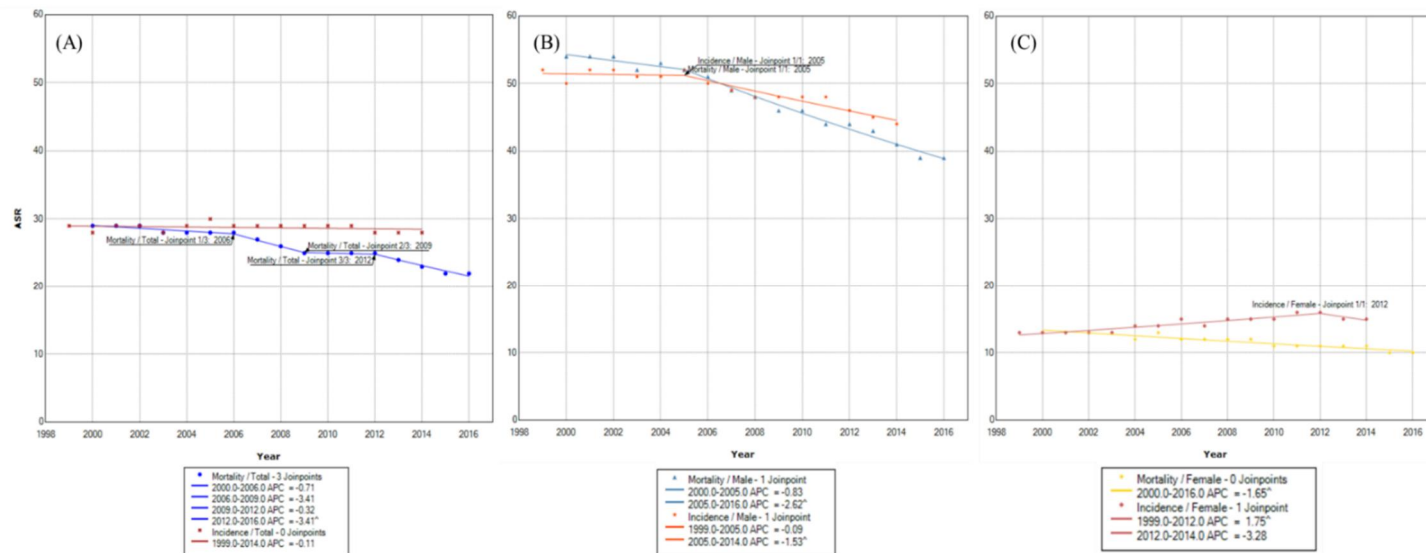


Figure I -2. Trends in age standardized rates in lung cancer incidence and mortality in Korea based on the joint point regression  
(A) Both sexes (B) Male (C) Female

\* Data from KOSIS

## **2. The importance of suggesting survival prediction model to cancer survivors**

As the number of cancer survivor increases, improving quality of cancer care has become a major concern for medical relatives and the government [2–4]. A high qualified cancer management program involves medically appropriate follow-up care as well as health information accessibility for self-management [5], which can lead to an effective shared decision-making process and improved medical outcomes [6]. Here, clinicians should concern with providing cancer survivors with an appropriate level of information in order to lower survivor's decision conflicts. [23]. As cancer can be life-threatening [22], and majority of the statuses are uncertain, the perception of inadequate information regarding health related problems may be critical to the internal conflict of patients.

Therefore, cancer survivors who overcome the immediate effects of cancer treatment need more evidence-based information that suggests monitoring of multidimensional health related problems, such as physical, psychological, social, and spiritual health issues.[16, 17] One of the major challenges in surviving lung cancer is to classify patients into exact prognostic group and provide them appropriate information for better follow-up planning and personalized self-management.[18] A variety of prediction models for lung cancer mortality have been developed and utilized in clinical setting, [19, 20] however, there were less studies developing survival prediction model based on the HRQOL factors. Providing appropriate information on assessing lung cancer mortality is critically important not only to improve patients' quality of life, but also to optimize patients' health and to help in patients' self-management to cope with lung cancer recurrence

or death.

### **3. HRQOL and lifestyle measurement as important predictors for lung cancer survival**

In addition to the traditional assessments of clinical outcomes, HRQOL or life style factors such as obesity or physical activity can play important roles in surviving long-term cancer. In fact, many lung cancer survivors reported that they suffered from diverse health difficulties [13, 17, 21] and their health function or symptom burden were severe in comparison to others, even if there were no clinically significant differences between the groups in terms of survival time [17, 21, 22]. Considering that many lung cancer survivors experienced physically impaired cardiorespiratory fitness [22-24], fatigue [17, 21-23], cachexia [25, 26], and appetite loss (anorexia) [17, 21], they tended to face worse prognoses in comparison to other cancer patients, and their HRQOL could be the important predictors for lung cancer survival.

In this respect, HRQOL data or patient-reported outcomes, which can be used as measures of the overall well-being and functioning of patients, may be utilized as complementary monitoring tools in routine follow-up practice for cancer survivors. While fixed clinico-pathological information is challenging to be modified, HRQOL factors or lifestyle factors can be calibrated by the health behavior modification. Routine assessment of HRQOL in oncology practice positively impacts physician–patient communication, and improves medical outcomes and emotional functioning in some patients [27]. Although earlier studies suggested that physical symptoms, such as anorexia [28], pain [28-31], and fatigue [28] are the strongest independent prognostic factors for cancer survival even after

the adjustment for established prognostic variables, mental health criteria, such as psychological distress, existential well-being, and post-traumatic growth, could also be independent predictive contributors for long-term survival among long-term cancer survivors [32-34]. In addition to their utility in assessing patient well-being and facilitating clinical decision-making, recent studies have suggested that HRQOL data can also provide distinct prognostic information [35, 36]. Global quality of life (QOL), functioning domains, and symptom scores were shown to be predictive of survival duration in various cancers, including breast cancer, lung cancer, and head and neck cancers [37-39].

Persuasive evidence indicates that obesity can also cause survival risk for cancer survivors; however, PA has a protective effect [40-44]. Although weight and PA guidelines for survivors should be tailored according to the type of cancer, the effect of obesity and PA on survival rates among lung cancer patients remains controversial [40]. Furthermore, there have neither been mortality studies reported for obesity and PA in lung cancer survivors, nor do we understand the effect of weight gain. Therefore, in this study, we extend our previous study [45] of lung cancer survivors through 5 more years of follow-up to address the risk factors of body mass index (BMI), PA, and HRQOL toward lung cancer mortality.

Therefore, it is of importance to predict cancer survivors' HRQOL or lifestyle factors in advance, and monitor their QOL and provide an appropriate education program. Even if there were a number of studies that investigated survivors' HRQOL as prognostic factors, there are limited studies that developed a cancer survival prediction model based on HRQOL factors or lifestyle factors [46, 47]. If HRQOL factors are independent predictors of survival in lung cancer, they

could be used in daily clinical practice to identify patients who will benefit from a specific intervention. Furthermore, it could help to set up more personalized psychosocial intervention programs aimed at improving patients' HRQOL [46].

#### **4. Traditional survival analysis versus machine learning techniques (MLTs)**

Predicting the time of survival accurately is a critical problem in longitudinal data analysis. For most of the real-world applications, the primary objective of monitoring these observations is to obtain a better estimate of the time of death. Here, traditional statistical models, such as Cox proportional hazards regression and some Kaplan-Meier models can be used to predict days till participants' death. Generally, survival analysis methods can be classified into two main categories: traditional statistical methods and machine learning based algorithms. In addition to statistical methods and MLTs possessing the common goal to predict survival, they both focus more on the distribution of time until the occurrence of the event. Survival analysis includes complex events, data transformation, and early prediction.[48] A complete classification of these survival analysis methods is shown in Figure I -3.

MLTs are generally applied to high-dimensional problems, while traditional statistical methods are generally developed for low-dimensional data. In addition, MLTs for survival analysis offer more effective algorithms by incorporating survival problems with both statistical methods and machine learning methods, taking advantages of the recent developments in machine learning. Even if the nature of machine learning algorithms which are referred to as the “black box



model”, machine learning-based decision supporting may have potential to health care domains showing high accuracy in prediction modeling.[19, 49] Therefore, machine learning algorithms, such as decision trees (DT), artificial neural networks (ANN), and support vector machines (SVM), which have become more popular in the recent years, are included under a separate branch. Several advanced machine learning methods, including ensemble learning (i.e., adaptive boosting (AdaBoost), random forest (RF), and bagging) are also included. Those of advanced MLTs can drive changes in health care, specifically in cancer prognostic models.

In general, statistical approaches focused on inferring the characteristics of a population from sample data, while machine learning is focused on predicting future values by analyzing given data. Therefore, machine learning will be used for prediction problems because it learns for the purpose of maximizing the prediction accuracy. In fact, machine learning algorithms contain a lot of statistical techniques unknowingly. In this viewpoint, the consideration of MLTs application of lung cancer prediction model can be meaningful.

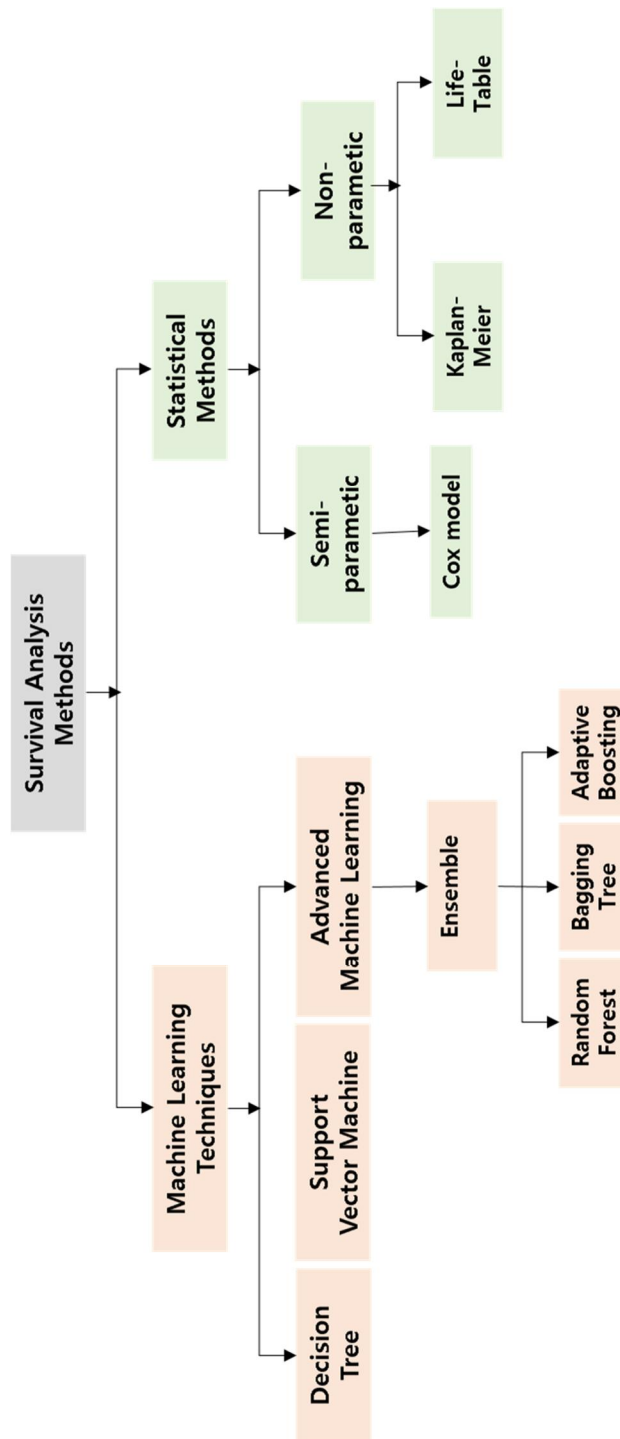


Figure I -3. Classification of survival analysis methods

## **B. Hypothesis and objectives**

### **1. Hypothesis**

Figure I-4 depicts the hypothetical diagram of the study on the development of lung cancer survival prediction model in Korea. This study aimed to comprehensively develop a valid model to predict lung cancer survivors' 5-year mortality using diverse MLTs and compare these models with that of the Cox regression survival model.

### **2. Objectives**

- I. To select the candidate prognostic factors of lung cancer mortality through literature review and to suggest the evidence for identification of the major factors that contribute to lung cancer survival.
- II. To identify the best mathematical model that explains individual prognostic factors of lung cancer survivors' traditional clinical variables and HRQOL measurements, and its interaction in the development of lung cancer survival prediction model based on the Cox regression survival model.
- III. To apply the prediction models into the MLTs and evaluate the validity of the developed mathematical model within data, and to establish the best model in comparison four MLTs (DT, SVM, RF, bagging, AdaBoost).

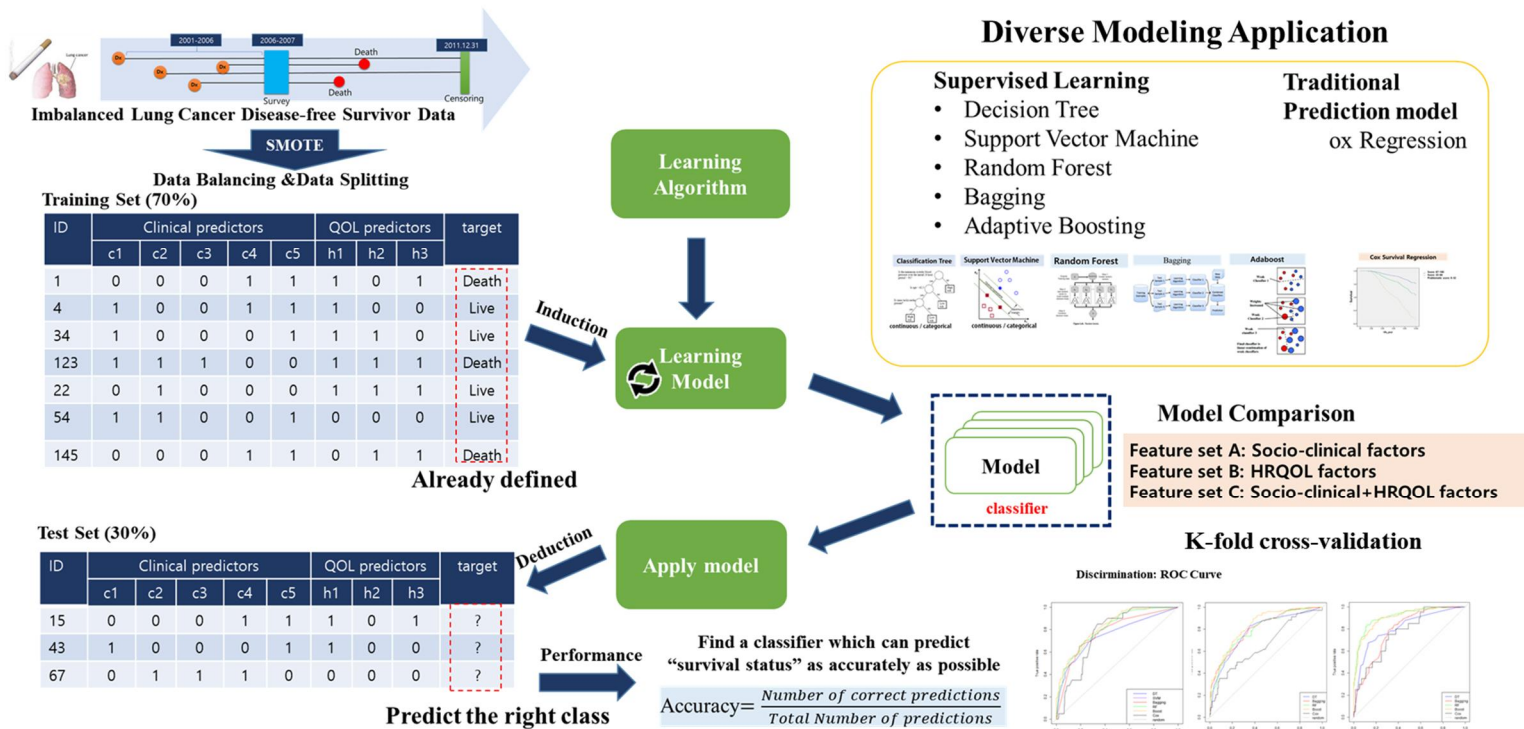


Figure I -4. Hypothetical series of tasks diagram of the current study

## **II. MATERIALS AND METHODS**

### **A. Study subjects**

#### **1. Subject selection**

The study participants consisted of 809 patients aged over 18 years who underwent lung cancer surgery between 2001 and 2006 at the Samsung Medical Center or the National Cancer Center in South Korea. The participants who were not disease-free for at least 1 year were also excluded. Disease-free survival time was defined as the time from the age or date of lung cancer surgery to the age or date of recurrence, including loco-regional recurrence, first distant metastasis, contralateral lung cancer, second primary cancer, and any cause of death.

Data was available on date of diagnosis, primary cancer site, disease stage, type of treatment, and other clinical characteristics. Information on health behavior and QOL was collected by professional interviewers who visited each patient's residence and administered the questionnaire. The patients were pathologically diagnosed as disease-free at the time of the study, and did not receive any treatment while the study was in progress.

The patients were eligible to participate if they (1) had a past diagnosis of lung cancer, (2) were treated with curative surgery, and (3) had no other history of cancer. Eligible subjects were contacted by telephone, and those who agreed to participate were surveyed by an interviewer with the help of questionnaires at home or the clinic. In this analysis, we excluded the subjects whose cancer had recurred at that time. As video-assisted thoracic surgery was not often performed

from 2001 to 2006, we also excluded patients who received it. Thus, all patients in this study underwent pulmonary resection through open thoracotomy.

Among such patients, we excluded 27 subjects whose survival status was censored until December 31, 2011. Thus, a total of 809 patients were included in this study. We collected information in relation to the date of the diagnosis, stage, type of treatment, and other clinical characteristics from the hospitals' cancer registries. This study was approved by the Institutional Review Boards of each hospital. The criteria for enrollment and study details have been elaborately described previously [17]. The whole process of study subject selection is shown in Figure II -1.

## **2. Data collection**

A standardized questionnaire was provided to trained interviewers to collect information on patients' socio-demographic factors, past medical history, lifestyle factors, and HRQOL factors. Information on health lifestyle and QOL was gathered by professional interviewers who visited each patient's residence by administering a semi-structured questionnaire. The questionnaire used for the cases and controls were identical. The specific composition of the questionnaire and the clinical pathology factors are listed in Table II -1.

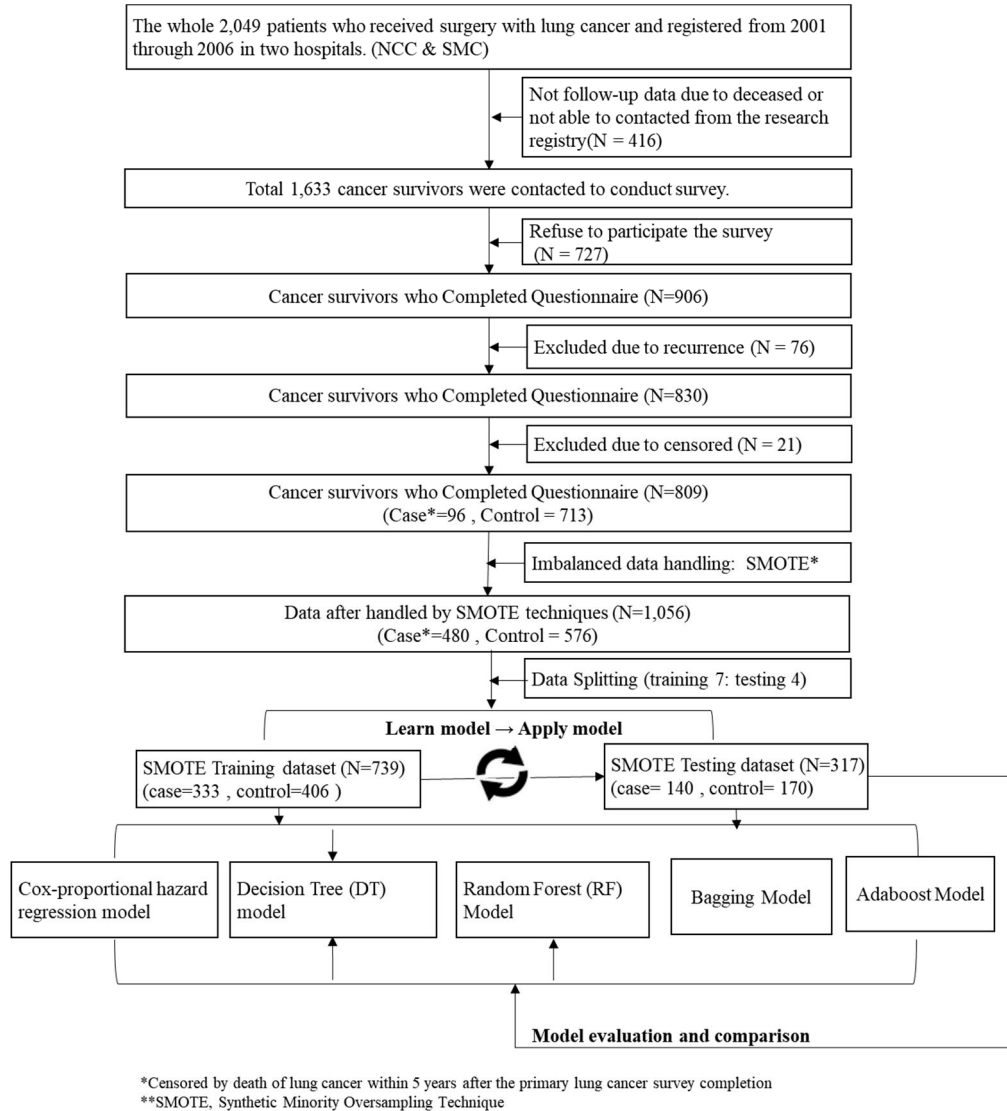


Figure II-1. Selection of eligible study subjects in the current study

Table II-1. Past medical history information and composition of questionnaire

---

**Categories**

---

**Past Medical History (data driven from cancer registry)**

Stage basis (pathological or clinical)

Local invasion of tumor

Regional lymph node metastasis

Type of treatment

FEV1/FVC ratio

Recurrence status

Number of comorbidities

**Patient Socio-demographics**

Age

Sex

Residence

Marriage

Education status

Monthly income

Job status

Years from survey date to diagnosis date

**Patient Lifestyle Characteristics**

Alcohol experience

Smoking experience

BMI at survey

Weight gain

Physical activity

**HRQOL Assessment**

EORTC QLQ-C30 (Functioning, Symptoms)

EORTC QLQ-LC13 (Functioning, Symptoms)

PTGI, Post-traumatic Growth Inventory

HADS, Hospital Anxiety and Depression Scale

---



### ***2.1. Socio-demographic and clinical variables***

We used a combination of published questionnaires to gather demographic (age and sex), socioeconomic (marital status, educational level, monthly family income, and place of residence), and clinical data (cancer stage, local invasion of tumor, regional lymph node metastasis, regional lymph node metastasis, FEV1/FVC (Forced expiratory volume in 1 second / Forced vital capacity), recurrence, number of comorbidity, treatment type, time since diagnosis, and years from survey date to diagnosis date). To identify the influence of comorbidities on cancer patients, we asked them about the current existence of comorbidities, such as cerebrovascular disease (e.g., stroke or cerebral hemorrhage), heart disease (e.g., angina pectoris, myocardial infarction, or chronic heart failure), diabetes, liver disease (e.g., chronic hepatitis or cirrhosis), pulmonary disease (e.g., chronic bronchitis or asthma), hypertension, infectious diseases (e.g., tuberculosis), digestive diseases (e.g., chronic gastritis, gastric ulcer, or duodenal ulcer), musculoskeletal disorders (e.g., degenerative or rheumatoid arthritis), and kidney disease (e.g., chronic renal failure).

## ***2.2. Patient lifestyle characteristics***

### **2.2.1. Body mass index and weight change**

To calculate the BMI, we obtained information on each patient's height and weight before surgery and after treatment from a self-administered questionnaire. For Asian populations, the most frequently employed BMI cut-off point for risk of type 2 diabetes and cardiovascular disease ranges from 23–25 kg/m<sup>2</sup>. We classified participants into two BMI categories: below 23.0 and overweight as a BMI > 23.0 kg/m<sup>2</sup> [50], and two weight-gain categories ( $\leq 4$  kg and  $> 4$  kg).

### **2.2.2. Physical activity**

We assessed leisure time PA in metabolic equivalents of task (MET)-hours per week during the past year. Patients were asked “During the past year, select the most regularly exercised physical activity, average time, and hours spent per week.” The options included the following 12 activities: walking, hiking, running (jogging), weight training, playing football, swimming, golfing, playing tennis, calisthenics, aerobic dancing, playing basketball, and bicycling. We calculated the total number of hours of PA from the frequency and average number of hours engaged in moderate activities (e.g., walking and calisthenics) to brisk activities (e.g., running and strenuous sports), and estimated the MET [8]. We considered PA to be at least 30 minutes of moderate-to-vigorous PA 5 or more days per week (i.e.,  $\geq 12.5$  MET/week).

### ***2.3. Health related quality of life (HRQOL)***

Patients completed questionnaires that covered the following demographic characteristics: the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core-30 item (EORTC QLQ-C30) and lung cancer module (QLQ-LC13), Hospital Anxiety and Depression Scale (HADS), and Post-traumatic Growth Inventory (PTGI).

#### ***2.3.1. EORTC QLQ-C30***

The EORTC QLQ-C30 is a 30-item cancer-specific questionnaire for measuring global health and overall QOL scales, five functioning domains (physical, role, cognitive, emotional, and social), three symptom scales (fatigue, pain, and nausea and vomiting), and six single items that assess additional symptoms commonly reported by cancer patients (dyspnea, appetite loss, sleep disturbance, constipation, and diarrhea), along with any perceived financial challenges [51].

#### ***2.3.2. EORTC QLQ-LC13***

The QLQ-LC13 incorporates one multi-item scale (dyspnea) and nine single items (pain in the arm/shoulder, chest, and other organs; cough; hemoptysis; dysphagia; peripheral neuropathy; alopecia; mouth sores). In both the surveys, high scores represent better functioning and severe symptoms. We dichotomized each scale of EORTC QLQ-C30 and EORTC QLQ-LC13 based on the score for the problematic group:  $\leq 33$  on a scale of 0-100 for globalQOL or functioning scale and  $> 66$  for symptom scale [52].

### ***2.3.3. Hospital Anxiety and Depression Scale (HADS)***

HADS is a self-reported assessment tool that comprises of two domains: the anxiety subscale and the depression subscale [53]. Each of the two HADS-subscales was measured using seven items rated on a 4-point Likert scale ranging from no feelings of anxiety or depression (0) to severe feelings of anxiety or depression (3). The total scores ranged from 0 to 21 for each anxiety and depression subscale. We used HADS as the outcome measure, which was dichotomized with the cut-off point of 8 as a borderline case of anxiety or depression [54].

### ***2.3.4. Post-traumatic Growth Inventory (PTGI)***

PTGI includes 21 items of positive changes, with five domains relating to others, personal strength, new possibilities, appreciation of life, and spiritual change. Each question was scored from 0 to 5 using a 6-point Likert scale. A higher score signifies greater post-traumatic positive growth [55]. We dichotomized each variable of PTGI according to the PTGI manual [55].

## 2.4. Definition of overall survival (OS) data

Overall survival (OS) time were defined as the time from the date of lung cancer surgery to the date of any cause of death. Lung cancer patients who did not have evidence of recurrence or death were censored in last follow-up until the target date. In this study, a regular follow-up was undertaken for the patients based on each hospital's registry after the completion of treatment. If the patients died during the follow-up, the family caregivers were asked the date of death. To obtain the date of death for the study subjects, we used the National Statistical Office database for dates through December 2009 and the hospital databases from then to December 31, 2011 (Figure II -2). We measured survival time from the date of the diagnosis and used mortality data with vital status. The person-years at risk data were accumulated for each patient from the date of the survey to the date of death. During the follow-up of 4509.2 person-years, we identified 96 deaths (11.9%) among the 809 subjects. In the 809 lung cancer survivors for whom there were available data, the median time from the diagnosis to survey date was 6.0 ( $\pm 1.24$ ) years and the median survival time was 8.3 ( $\pm 2.01$ ) years.

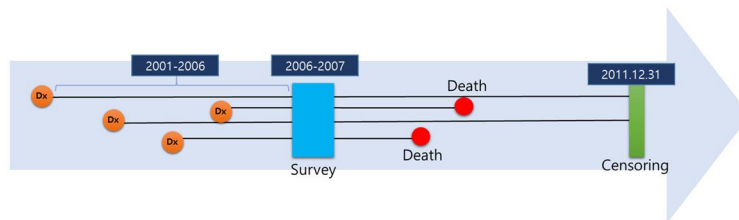


Figure II-2. Overall survival (OS) data structure

### 3. Study process

To develop a survival prediction model using Korean cancer survivor's HRQOL cross-sectional data, we followed the four steps mentioned below (Figure II-3). In this paper, we only focus on Step 1 to Step 3.

Step 1: Examination of the variables, data preprocessing

Step 2: Prediction model development

Step 3: Prediction model validation

Step 4: Application to clinical and medical setting

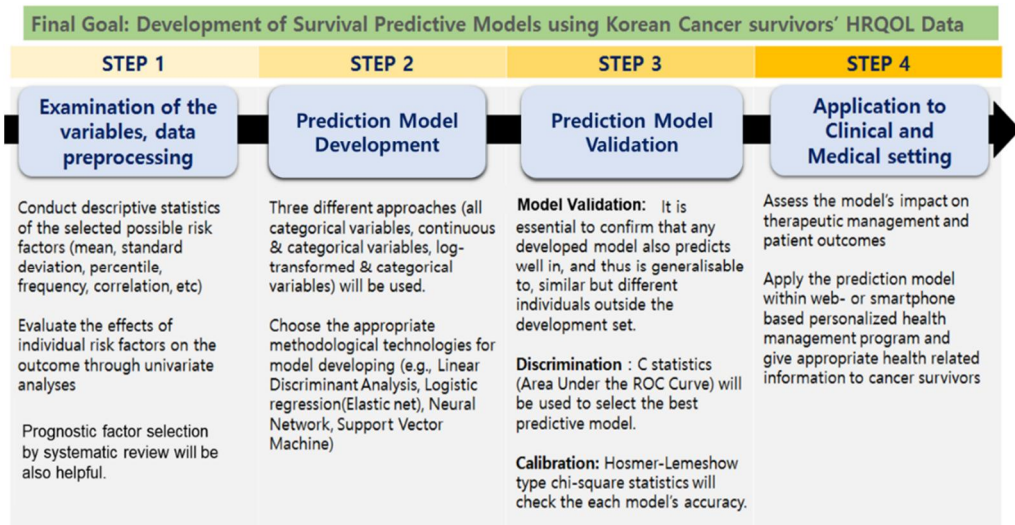


Figure II-3. Prediction model development and validation process

## ***B. Prognostic variables' selection and data preprocessing***

### **1. Prognostic variables' selection**

#### **1.1. Literature review for the selection of candidate predictors**

In the process of prediction modeling, and in order to increase the robustness and validity of a model, selection of candidate predictors is important. Ideally, the candidate predictors were chosen before studying the predictor-outcome relationship for the data under study. Considering further understanding of the prognostic factor analysis of HRQOL and lifestyle measurement data in cancer, a systematic review is required. Therefore, to select candidate predictors, a systematic literature review was conducted. The main research question for the literature review was “What are the relevant candidate HRQOL and lifestyle factors for lung cancer survivors?”

In order to conduct this systematic review, a lung cancer survivor was defined as any individual who was diagnosed with lung cancer and completed all their treatment until the end of their life. Long-term lung cancer survivors have been defined as people who live more than 5 years after diagnosis with or without disease [56, 57]. HRQOL and lifestyle factors were defined as the QOL related to one's health and functioning status, assessing the subjective perception of an individual's function or symptoms in the context of the individual's health condition and personal factors [58-60]. A candidate predictor was defined as any traditional well-known prognostic factors (i.e., related to demographic, clinico-pathological, and social characteristics) having an association with HRQOL, either cross-sectional or longitudinal studies.

Prior to the further consideration of prognostic variables, patient intervention comparison outcome (PICO) strategy for literature reviews related to lung cancer survival prognostic studies was adopted. The PICO strategy questions are described in Table II -2. An online literature search was conducted using 'Pubmed' and 'Medline', where the search terms included "lung cancer" or "lung neoplasms", "prognostic factors" or "survival", and "HRQOL" or "BMI," "Weight," "Obesity," "Physical Activity," or "Exercise." The search was restricted to English language and human studies, while search date was restricted from January, 2000 to September, 2017.

In addition to publication titles, the literature was examined to ensure that the study used a HRQOL instrument or measured HRQOL using validated indicators, and applied multivariate analyses for lung cancer survival adjusted for one or more well-known clinical prognostic indicators. Purely psychological studies or studies investigating predictors associated with HRQOLs and lifestyle factors were excluded. These were defined as studies limited to the relationship between one or more psychological variables, such as fighting spirit, cancer personality, coping styles, hostility, etc. and survival duration. Meta-analysis articles investigating the clinical, socio-demographic, and epidemiologic factors associated with progression or risk factors of the lung cancer survivors.



Table II-2. Patient intervention comparison outcome strategy questions

Factor	Research Questions	Contents
Patient, Population	What are the characteristics of the patient or population?	Lung cancer survivor Disease-free lung cancer survivor
Intervention or exposure	Generally time or “watchful waiting.”	More than 3-year disease free survival
Comparison ( if appropriate)	Generally not applicable for prognosis factor review	NA
Outcome	What you are trying to accomplish, measure?	Overall survival (OS) rates, mortality rates. What is the hazard ratio of death?
Study design	What are the study designs of the searching papers?	Questions of lung cancer survival or likelihood of a death

## **1.2. Grading the evidence and mapping into the conceptual framework**

During data extraction, potentially relevant candidate predictors, including HRQOL and clinical or socio-demographic variables considered together were initially identified based on the literature review. Further, the strength of evidence for identified prognostic factors was assessed by stepwise scoring and grading procedure, based on previously recommended procedures that consisted of three consecutive steps [61]. Stepwise scoring and grading procedure applied to assess the quality of evidence are shown in Figure II -4.

First, a quality of score was assessed according to each individual study based on the study design (longitudinal study versus cross-sectional study) and sample size ( $n < 100$  versus  $n \geq 100$ ). As previous studies suggested [61-63], the methodological rationale for quality score assessment was that longitudinal studies could provide more valid and stronger evidence in comparison to cross-sectional studies, and that larger sample size studies provide more reliable evidence in comparison to smaller sample size studies [63].

Second, consistency of evidence for each of the identified prognostic factors was investigated across different studies by summing the quality scores assessed from “step 1” for individual studies that observed the same association of a particular prognostic factor. An established World Health Organization’s International Classification of Functioning, Disability, and Health (ICF) linking procedure was applied to group factors that were conceptually similar by linking them to the corresponding ICF category.[58, 64, 65] The prognostic factors were

grouped into body Function and structure, activities, health condition, environmental factors, and personal factors. The consistency of evidence was categorized as follows: Category A: highly consistent (sum of scores from step 1:  $\geq 6$  points), Category B: moderately consistent (sum of scores from step 1: 4–5 points), and Category C weakly consistent (sum of scores from step 1:  $< 4$  points). [61]

The final step graded the total strength of evidence for the potential candidate of prognostic factor. For each identified prognostic factor, the evidence was graded as “strong” if it was a factor rated as “category A” from step 2; “moderately inconsistent” if it was rated as “categories B or C” from step 2; or “inconsistent” in case of contrary findings in different studies from step 2.

Assign each of quality score according to individual lung cancer HRQOL prognostic factor studies



- **Score 0**  
Cross-sectional study, sample size (N<100)
- **Score 1**  
Cross-sectional study, sample size (N<100)
- **Score 2**  
Cross-sectional study, sample size (N<100)
- **Score 3**  
Cross-sectional study, sample size (N<100)



Sum of scores from all articles observing the HRQOL prognostic factors of lung cancer survival



- **Category A (Highly consistent)**  
Sum of scores from step 1: 4-5 points
- **Category B (moderately consistent)**  
Sum of scores from step 1:  $\geq 6$  points
- **Category C (weakly consistent)**  
Sum of scores from step 1:  $<4$  points



Grade all review evidence determining candidate prognostic factors of lung cancer survival



- **Strongest evidence**  
Category A from step 2
- **Moderately inconsistent**  
Categories B or C from step 2
- **Inconsistent**  
Contrary findings in different studies from step 2

Figure II-4. Stepwise scoring and grading procedure applied to assess the quality of evidence [61]

### **1.3. Examination of prognosis variables' selection from statistical analyses**

After conducting a literature review to select candidate prognostic variables, univariate analyses were conducted to select relevant variables as independent variables in the lung cancer survival prediction model. In this process, descriptive statistics of the selected possible prognostic factors suggesting the mean (standard deviation (SD)), percentile, and frequencies were analyzed between groups of dead or alive. Although the use of established literature or knowledge is one of the most representative methods for preliminary screening of affective independent variable selection [18], this would introduce a significant bias in the variable selection process. Therefore, univariate analyses for categorical or continuous variables and univariates were applied additionally. Variables which were selected from the literature review suggesting at least weak evidence, and also showed the significance from statistical analyses were finally selected as candidate variables to develop the model.

## 2. Data preprocessing

### 2.1. Data cleaning, missing imputation

Data quality is a major concern in machine learning. As most machine learning algorithms strictly induce knowledge from data, the quality of the knowledge extracted is largely determined by the quality of the underlying data [66]. Therefore, missing values imputation for compositional data using classical and robust methods are preferable. Although there are several methodologies to treat missing values, in this part, we attempted to apply the imputation methods which involve replacing missing values with estimated ones based on some information available in the data set. Among them, we used “k-nearest neighbor (KNN) algorithm” to estimate and substitute our missing data. KNN algorithm is useful as it can predict both binary and continuous features. Using R packages of “DMwR,” we can replace the weighted average numbers of the nearest neighbors with missing values. In our data, we apply 5 neighbors ( $k=5$ ) in our algorithm.

### 2.2. Test of multi-collinearity

Broadly, highly correlated factors in prediction models have the following implications:

- They can increase the standard error (SE) of estimates of the  $\beta$  coefficients.
- They can often lead to confusing and misleading results.
- If the interest is only in estimation and prediction, high correlation can be ignored as it does not affect  $\hat{y}$  or its SE (neither  $\hat{\sigma}_{y\wedge}$  nor  $\hat{\sigma}_{y-y\wedge}$ ).

As the SEs of estimated  $\beta$  coefficients are higher with highly correlated factors, which can create less certainty about the developed predictions,

certainly, they can be more challenging to explain. Therefore, the criteria to identify highly correlated variables were high correlation coefficient values larger than 0.7 [67].

$$\rho = \frac{cov(X, Y)}{\sigma_x \sigma_{xy}}, -1 \leq \rho \leq 1$$

### ***2.3. Decisions of cut-off points***

We considered PA to be at least 30 minutes of moderate-to-vigorous PA 5 or more days per week (i.e.,  $\geq 12.5$  MET/week and overweight as a BMI  $> 23.5$ . To maximize differences in prognostic strength of QOL scores, we dichotomized each variable score and chose a cut-off point. We dichotomized each scale of EORTC QLQ-C30 and EORTC QLQ-LC13 based on the score for the problematic group:  $\leq 33$  on a scale of 0–100 for global QOL or functioning scale, and  $> 66$  for symptom scale [52]. Earlier studies of cancer survivors have shown that the scores for the problematic group were useful in identifying the problems of QOL in comparison to the general population [68, 69].

## 2.4. Data sampling for data balancing, SMOTE

If the classes of dataset are not approximately and equally represented, data can be imbalanced and can be prevalent in fraud detection [70]. The performance of machine learning algorithms is typically evaluated using predictive accuracy, if the data is imbalanced and the costs of different errors vary markedly.[71] In this situation, to reduce the error cost, data balancing methods such as “over-sampling” or “down-sampling” may be useful, while over-sampling or down-sampling with replacement does not significantly improve minority class recognition. Therefore, we used the synthetic minority over-sampling technique (SMOTE), which is a type of over-sampling approach in which the minority class is over-sampled by creating “synthetic” examples and joining any of the  $k$  minority class nearest neighbors rather than merely by over-sampling the replacement. In our code, we balanced “dead” and “alive” from our data by over-sampling 500 “dead” sets, and under-sampling 100 “alive” sets. We used the R studio packages of “DMwR.” Comparison of balancing methods including original, down-sampling, up-sampling, and hybrid smote are shown in Figure II-5.

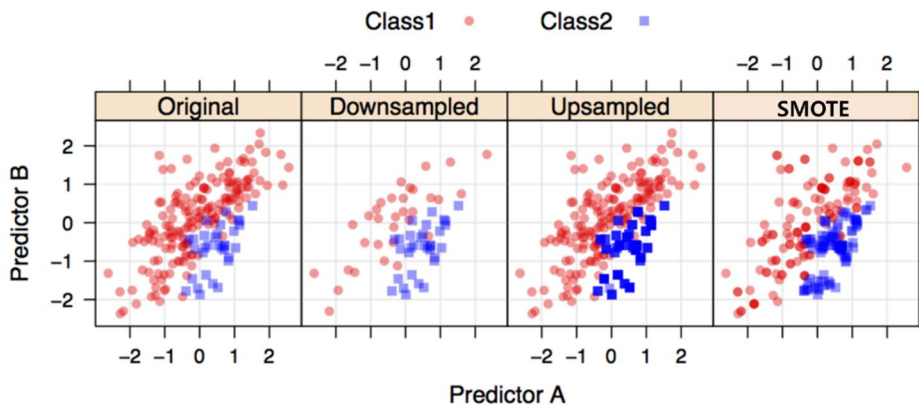


Figure II-5. Comparison of balancing methods



## 2.5. Data splitting (holdout strategy)

The holdout method in data splitting was employed to avoid over-fitting in the model and to derive reliable estimates of the model performance. Model over-fitting could arise when the number of events is small in comparison to the number of predictors in the predictive model. The holdout method randomly splits the whole data sample into two mutually exclusive training (70%) and testing (30%) sets. The training set was utilized to generate the prediction model and the remaining 30% of the data was employed as a testing set to estimate the model's accuracy. Process of hold-out sampling method to avoid over-fitting problems are shown in Figure II-6.

d

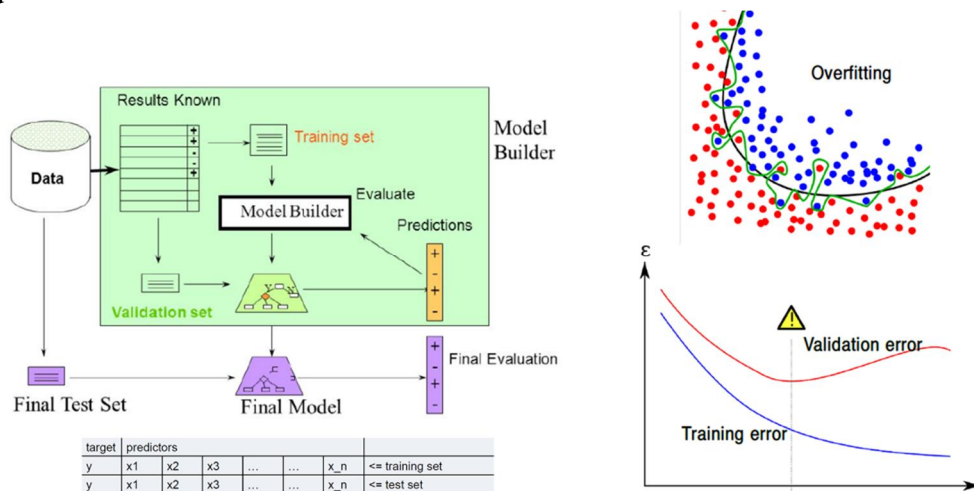


Figure II-6. Hold-out sampling method to avoid over-fitting problems

## **C. Model development**

The progress to surviving lung cancer, similar to many other chronic diseases, is seldom the result of a single cause, rather than the result of a complex combination of clinical and HRQOL problems and exposure to various risk factors throughout a person's life. With advances in statistical methodology, prognostic appraisal tools are increasingly being developed and used to estimate individual prognostic factors of developing or dying from cancer. Statistical modeling can be categorized into linear modeling and nonlinear modeling, depending on the predictor functions used for modeling and are described briefly in the following section.

Although many data mining algorithms have been developed, this study entailed the development of the traditional Cox proportional hazard regression model and application toward four MLTs, including 1) DT, 2) RF model, 3) Bagging, and 4) AdaBoost, which were used to find the best model that describes data. Each of modeling techniques were applied according to the feature set 1, 2, 3, and those of models were compared among their model performance within the same modeling techniques.

## 1. Cox model development

The Cox proportional hazard regression models were used for developing lung cancer survival prediction equations in the lung cancer survivor development set. The subjects were censored at the date of death ascertained from the death certificate database from the Korean Statistical Information Service, or on the end date after 8 years of follow-up.

The baseline survival estimate for the mean values of the risk factors for time  $t$  ( $t = 5$  years) was estimated by the following equations:

- $S(t, X) = [S_0(t)]^{\exp(\sum_{i=1}^k \beta_i (X_i - M_i))}$
- $P(\text{event}) = 1 - S(t, X).$

Here,  $\beta_1 \sim \beta_k$  are the regression coefficient estimates,  $x_i \sim x_k$  are the risk factors for each individual, and  $M_1 \sim M_k$  are the mean values for each prognostic factor among the participants.  $S(t)$  is the baseline survival estimate at time  $t$  ( $t = 5$  years) when all the prognostic factors are at their mean values.

Considering statistical analysis, the dependent variable was dichotomized to be alive or death (event). The independent variables were entered as either binary or categorical. Crude and age-adjusted analyses were performed for each prognostic factor. The prognostic factors considered for the models were age, sex, FEV1/FVC ratio, stage, income, BMI, MET, and several QOL factors. All the risk factors were included as categorical variables in the model.

To select the model subject, stepwise-AIC best subsets approach in Cox regression were conducted. In the stepwise selection model, the sequence of models starting with the null model and ending with the full model (all the

explanatory variables included) is derived. The models in this sequence are ranked in order to maximize the increment in likelihood at every step. It is obvious to call this sequence the stepwise sequence. However, in our study, we attempted to find the minimum AIC. Model building in PROC PHREG from SAS 9.4 with automatic variable selection, constructing a full stepwise sequence, and shopping around optimal AIC were undertaken sequentially. Lower values indicated a better fit. Each of the Cox model from feature set 1, 2, 3 were developed. We will call them model COX-1, COX-2, COX-3 in order.

## **2. Decision tree (DT) model**

DT model can be used in a wide area of MLTs, covering both classification and regression. As our goal is to predict the binary target and classify patients in the correct event group, we attempted to develop our first machine learning algorithm with decision analysis, which can be used to visually and explicitly represent decisions and decision-making. As the name suggests, it uses a tree-like model of decisions. A DT is drawn upside down with its root at the top. The end of the branch that does not split anymore is the decision, whether the participants died or survived. In our study model, the DT starts splitting by considering each feature in the training data. The mean of responses of the training data inputs of a particular group is considered as a prediction for that group.

Preprocessing for the DTs are as follows: 1) undertaking missing imputation, excluding duplicated data, 2) holding-out sampling, and 3) identifying categorical variables and converting them to “factor” variables. As we use this algorithm in supervised learning classification, variables that are selected should be changed into categorical variables.

After the completion of preprocessing, DT predictive model development starts from variable selection and assesses the relative importance of variables. Many variables from the whole data set are of marginal relevance, and thus should probably not be included in the data mining process. Similar to the stepwise variable selection in regression model analysis, DT methods can be used to select the most relevant input variables that should be adapted to form DT models, which can continuously be used to formulate a clinical prognostic factor model. After a set of associated variables is identified, variable importance can be computed based

on the purities of nodes in the tree when the variable is removed [72].

The main components of a DT model are nodes and branches, and the most important steps in building a model are splitting, stopping, and pruning [73]. For splitting, characteristics that are related to the degree of “purity” of the resultant child nodes (i.e., proportion with the target condition) are used to choose between different potential input variables; these characteristics include entropy, Gini index, classification error, and information gain [72, 74]. A well classified model will show higher information gain. This splitting procedure continues until stopping criteria are met.

- Gini index:  $Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2$
- Entropy:  $Entropy(t) = -\sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t)$
- Classification error rate:  $Classification\ error(t) = 1 - \max[p(i|t)]$
- Information gain =  $Entropy(p) - (\sum_{i=1}^k \frac{n_i}{n} Entropy(i))$

The next step in the DT algorithm is pruning. One way to select the best predictable sub-tree is to consider the proportion of records with error prediction and the other is to use a hold-out data splitting using the training set, or for small samples, k-fold cross-validation to avoid over-fitting. Pruning is one of the MLTs which can reduce the size of a DT and prevent over-fitting of model training. Pruning is achieved by removing the nodes that have the least effect on the overall classification performance [75]. We used a training set to prune, pruned at a point which improves the accuracy of the overall classification and increases validation error when applied to the validation dataset. Each pruning step of model “cp” can be calculated and plotted as a figure. Finally, three DT models from feature set 1, 2, 3 were developed. We will call them model DT-1, DT-2, DT-3 in order.

### 3. Random forest model

The RF for survival analysis using prediction error curves was also evaluated for each model. An RF is a nonparametric MLT that can be used to build a prediction model in survival analysis. Within the survival settings, the prediction model was conducted by an ensemble learning formed by combining the results of many survival DTs [76].

The model development is based on RF as follows. First,  $B$  bootstrap survival tree is drawn based on the data of each of the bootstrap samples  $b = 1, \dots, B$ . Further, for each of the bootstrap samples, an unpruned classification tree is grown, rather than choosing the best split among all the predictors, “mtry” sample of the predictors is randomly explored and the best split is chosen. Finally, new data is predicted by aggregating the predictions of the  $B$  bootstrap survival trees. After obtaining the “error rate” of estimations based on the training data the error rate was calculated, which is called the out-of-bag (OOB) estimate of error rate. The OOB estimate is then aggregated [77]. The point of error rate remaining below the minimal rate provides the best fitted tree numbers.

Similar to the DT model, RF can also show the variable importance measures which are useful for model reduction. To build a simple model, the variable importance information provided are more readily interpretable models. The gain can be more dramatic when there are more predictors. For simulated lung cancer survival data and RF with a default “mtry,” we were able to clearly identify the best informative variables and totally ignore the other noise variables. Over-

fitting was controlled by OOB validation at 70% of the samples in comparison to DT pruning [78].

The R packages for RF is consistent to that of other classification functions, such as nnet package and svm ( ) from e1071 package. Considering this classification prediction, we specified the “factors” and used supervised learning. The function of RF returns an object using “randomForest” and “MASS” package. Finally, the methods predicted the right class and printed the results based on the test set. From the all process, each of the RF model from feature set 1, 2, 3 were developed. We will call them model RF-1, RF-2, RF-3 in order.

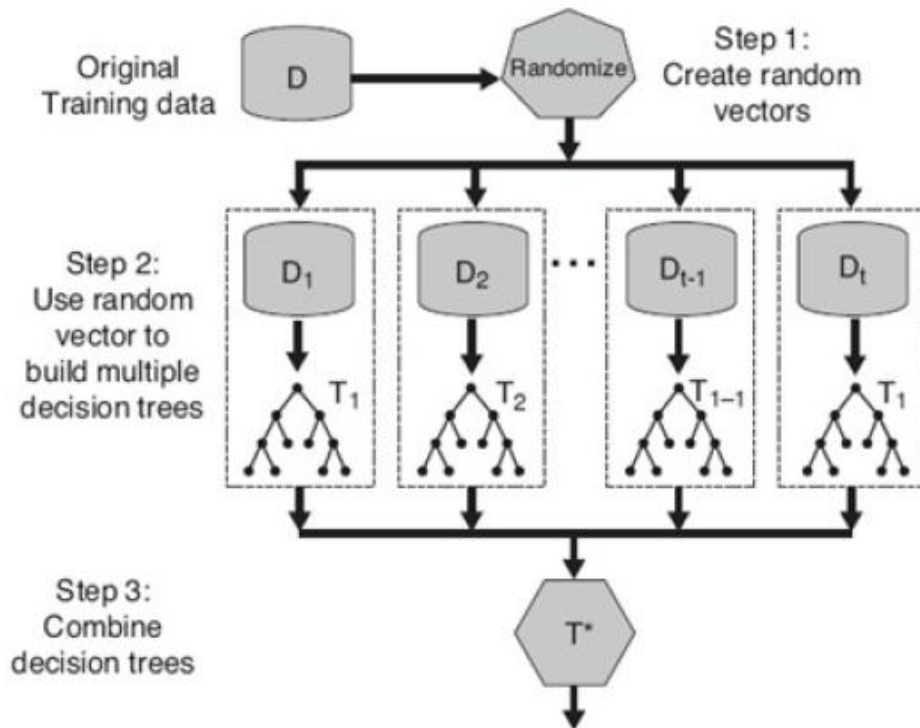


Figure II-7. Random forest algorithms



#### **4. Bagging (bootstrap aggregating)**

Due to the instability of tree-based DT modeling, the development of so-called ensemble methods, such as bagging or RF technologies which can lead to improvement of the predictability of several unstable classification methods are suggested. Considering this reason and for efficiency, the individual DTs are grown deep and the trees are not pruned. These trees will have both high variance and low bias. These are important characteristics of sub-models when combining predictions using bagging. After creating multiple bootstrap samples, multiple prediction model training for each sample set is developed, and finally the results of each model used to predict are combined [79].

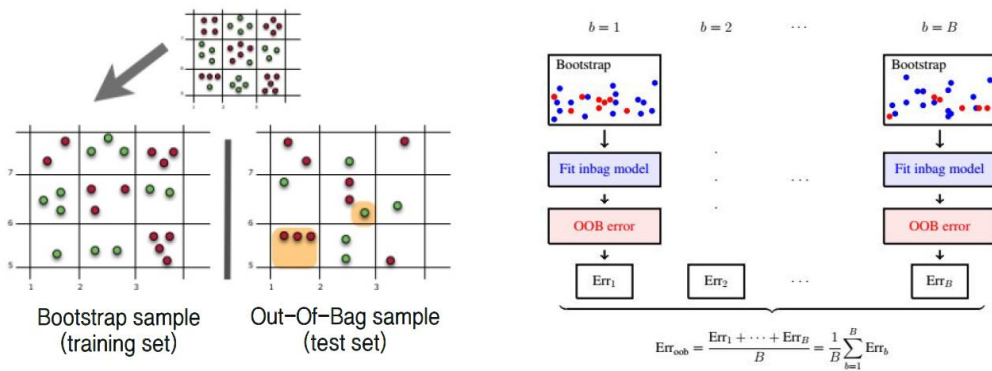
Similar to RF, the general strategy of bagging is to draw bootstrap samples from the original observations and to grow the maximal tree “mtry” for each of the samples. This strategy also circumstanced the problem of pruning and variable selection as each tree is grown to full size. The final prediction is obtained by averaging the predictions from each individual tree. In practice, bagging can be viewed as a special case of RF model where all the covariates are considered as relevant candidates at each node. These methods also provide a mechanism to define various variable importance measures, where basket selection can be used for variable selection.

Bagging is conducted in two ways, bootstrap sampling and aggregation. Bootstrap sampling is a technique to resample with replacement and extract the set of training data from the given learning data. On an average, the bootstrap sample contains 63.2% of the original data and this can be training data. In addition, data not extracted from the bootstrap sample, approximately 36.8%, are considered to

be test data. Considering the classification model, voting methods for categorical variables are use in aggregation.

$$P(\text{observation } i \in \text{bootstrap sample } D_i) = 1 - \left(1 - \frac{1}{N}\right)^N \approx 1 - e^{-1} = 0.632$$

The trees in this function are computed using the implementation in the “rpart” package. The generic function “ipredbagg” implements methods for different responses. As our target variable “y” is a survival object, bagging survival trees is performed. There is no general rule stating when the tree should be stopped from growing. By default, classification trees are as large as possible, whereas regression trees for bagging and survival trees are built with the standard options of rpart.control. For each of the models, the OOB sample is used to estimate the prediction error corresponding to the target event. Further, the final step is the model predicting the test set. From the all process, each of the bagging model from feature set 1, 2, 3 were developed. We will call them model Bag-1, Bag-2, Bag-3 in order.



## 5. Adaptive boosting (AdaBoost)

Boosting methods were originally proposed as one of ensemble methods, which rely on the principle of generating multiple predictions and majority voting (averaging) among individual classifiers. AdaBoost is an MLT that combines multiple weak learning algorithms to create a good classification model. Learning the classifier sequentially improves learning in the direction of complementing the disadvantages of the previous classifiers. AdaBoost is used to adaptively change the distribution of training samples, such that the default classifier focuses on challenging cases to classify [80].

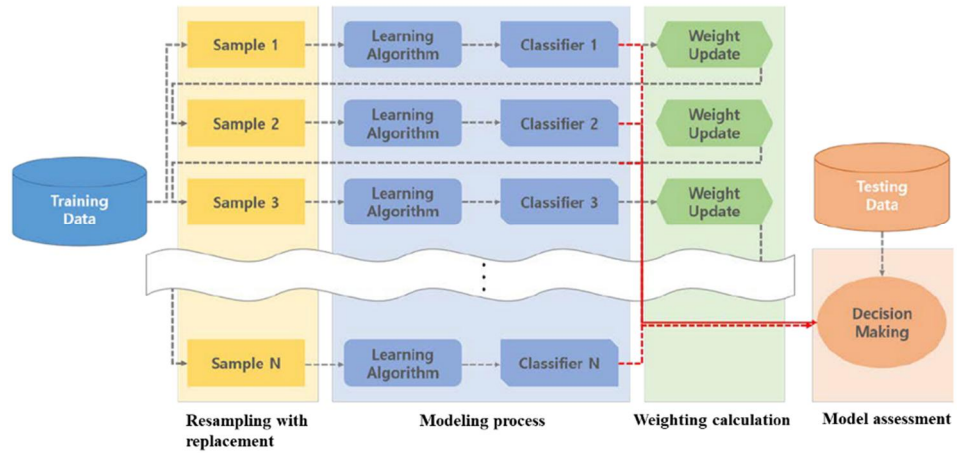


Figure II-9. AdaBoost model algorithms

Considering AdaBoost, learning the first classifier using initial data  $D_1(i)$  is undertaken. Re-distribution of the sample data by assigning a high weight to the sample data misclassified in the existing classifier is then followed. The aim is to select  $h_t$  with a low weighted error [80].

$$\text{Initialize: } D_1(i) = \frac{1}{m} \text{ for } i = 1, \dots, m$$

$$\text{Weight: } \alpha_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$

Finally, the classifier is learnt using the updated sample data repeatedly. Considering T times repeated learning of the classifier using the updated sample, the final strongest classifier is as follows:

$$H(x) = \text{sign}\left(\sum_{i=1}^T \alpha_t h_t(x)\right)$$

We can also plot the ensemble error according to the number of trees and find an appropriate number of trees for the training data. Similar to other ensemble learning algorithms, in AdaBoost, we can obtain information on variables' importance. Using the formula of "importance plot (boost)" in R, we can plot a variable's relative importance. The final step of the model prediction is using test set and evaluating model performance. Therefore, each of the AdaBoost model from feature set 1, 2, 3 were developed. We will call them model AB-1, AB-2, AB-3 in order.

## **D. Model validation**

### **1. Model validation for Cox model**

#### **1.1. Discrimination for Cox model**

The C-statistic is a concordance measure analogous to the receiver operating characteristic (ROC) curve area for the survival analyses model [16]. This value indicates the probability that a model produces higher risk for those who will die from lung cancer within 5 years of follow-up in comparison to those who do live [16]. Each model's discriminatory ability was tested by calculating the area under the curve (AUC) and the 95% confidence interval of the AUC. The difference between the models was tested by comparing the AUC values by Mann-Whitney U test. Both, performance of validation and test set, were analyzed.

#### ***1.2. Calibration for Cox model***

A Hosmer-Lemeshow (H-L) type  $\chi^2$  statistic was used for calibration [15]. The  $\chi^2$  statistic was calculated by first dividing the data into 10 groups (deciles) in ascending order of predicted probabilities produced by the model. Further, in each decile, the average predicted probabilities were compared to the actual event rate estimated by the Kaplan-Meier approach. Values exceeding 20 are considered to have a significant lack of calibration. In addition, calibration was tested, the expected number of lung cancer survivals ( $E$ ) was computed, and these were compared to the corresponding observed number ( $O$ ) of overall lung cancer survivals. The expected number of cases was calculated by summing the estimated individual absolute prognosis for each person predicted by the developed model.

Considering the Cox model performance assessment, all statistical analyses were performed using SAS version 9.4 (SAS Institute, Cary, NC)

## 2. Model validation of other MLTs

Among the randomly split hold-out data sample, the remaining 30% of the data (testing set) was employed to estimate the model performance, including the model's accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 measure and AUC with 95% confidence interval (CI). Each MLT based model was calculated for performance comparison between the proposed models. Each of the 12 models (applied with 4 MLTs, to predict with 3 feature sets)' ROC plots is suggested in addition to Cox models. All the comparisons and plotting were analyzed by R package (R Development Team, 2017) for all statistical analyses.

Training Confusion Matrix			Test Confusion Matrix			
Actual	Predict		Predict			
	Live	Death	Live	Death		
	Live	TN	FN	Live	TN	FN
	Death	FP	TP	Death	FP	TP

Figure II-10. Confusion matrices for the training dataset (left) and test samples (right).

The squares provide the performance metrics described in the following section.

**Abbreviation:** TP, True Positive; FP, False Positive; TN, True Negative; FN, False Negative

- $PPV \text{ (precision)} = \frac{TP}{TP+FP}$
- $NPV = \frac{TN}{TN+FN}$
- $Sensitivity \text{ (recall, TP rate)} = \frac{TP}{TP+FP}$

- Specificity (TN rate) =  $\frac{TN}{TN+FP}$
- AUC: ROC curve depicts sensitivity versus specificity at diverse discrimination thresholds and is commonly used in medical statistics [78].

### **3. K-fold Cross Validation for MLT based prediction models to avoid over-fitting**

Over-fitting is the phenomenon in which the learning system tightly fits the given training data so much that it would be inaccurate in predicting the outcomes of the untrained data. In decision trees, over-fitting occurs when the tree is designed so as to perfectly fit all samples in the training data set. In short, a decision tree is over-fitted if it gives highly accurate output on training data, but low accurate output on test data. Thus it ends up with branches with strict rules of sparse data. Thus this effects the accuracy when predicting samples that are not part of the training set.

Therefore, we used several approach to address the over-fitting for each of the MLTs. First, the methods to address over-fitting in decision tree, which is called pruning which is done after the initial training completes. In pruning, we can trim off the branches of the tree, for example, removing the decision nodes starting from the leaf node that the overall accuracy is not disturbed. This is done by segregating the actual training set into two sets: training data set, and validation data set. We call this hold-out sampling. Preparing the decision tree using the segregated training data set. Then continue trimming the tree accordingly to optimize the accuracy of the validation data set can also be helpful to avoid over fitting.

A better procedure to avoid over-fitting is to sequester a proportion (10%, 20%, 50%) of the original data, fit the remainder with a given order of decision tree, and then test this fit against the sequestered data. Overfitting is detected when the  $R^2$  for the sequestered data starts to fall below that fitted for the remainder. Some statistical r packages make it easy by using equivalent  $k$ -fold cross-validation ( $k=10,5,2$ ). In here, we used 5-fold cross-validation as following Figure II -11.

Relative to other models, Random Forests, or other ensemble techniques are less likely to overfit, but it is still something that we want to make an explicit effort to avoid. Tuning model parameters is definitely one element of avoiding overfitting but it isn't the only one. In fact, training features are more likely to lead to overfitting than model parameters, especially with an ensemble learning. Therefore, having a reliable method to evaluate the developed model to check for overfitting more than anything else. Even if choosing the best model based on  $k$ -fold cross-validation results will lead to a model that hasn't more overfit, which isn't necessarily the case for something like out of the bag error, we conducted the cross-validation. The easiest way to run  $k$ -fold cross-validation in R is with the caret package. In this thesis paper, we only showed the result of cross-validation in Appendix Table.



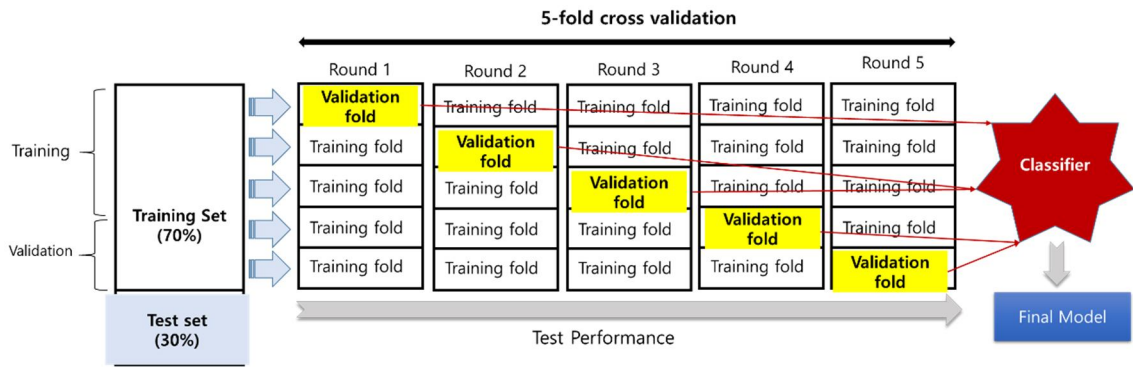


Figure II-11.  $k$ -fold cross validation ( $k=5$ ).

# III. RESULTS

## A. Literature review for selection of candidate predictors

### 1. Selection of candidate prognostic factors with literature review

A total of 460 articles were searched by the search term (“lung cancer” or “lung neoplasms”) and (“prognostic factors” or “survival”) in PubMed. Combined with the 34 articles out of the 426 articles and other articles found otherwise, finally, total of 25 articles were reviewed. (The flow of selection of candidate prognostic factors were shown in Figure III-1). Through citation tracking, lung cancer survival prognostic factors considered together with HRQOL factors or life-style factors were grouped into 5 ICF domain categories. The summary of schematic diagram of candidate prognostic factors from literature review mapped with on the ICF is provided in Figure III-2.

After excluding insignificant systematic review results, the total factors were grouped and assessed for the quality of scores (see Appendix). According to the evaluation criteria aforementioned, evidence of prognostic factors was ranked into highly consistent, moderately consistent, and inconsistent levels.

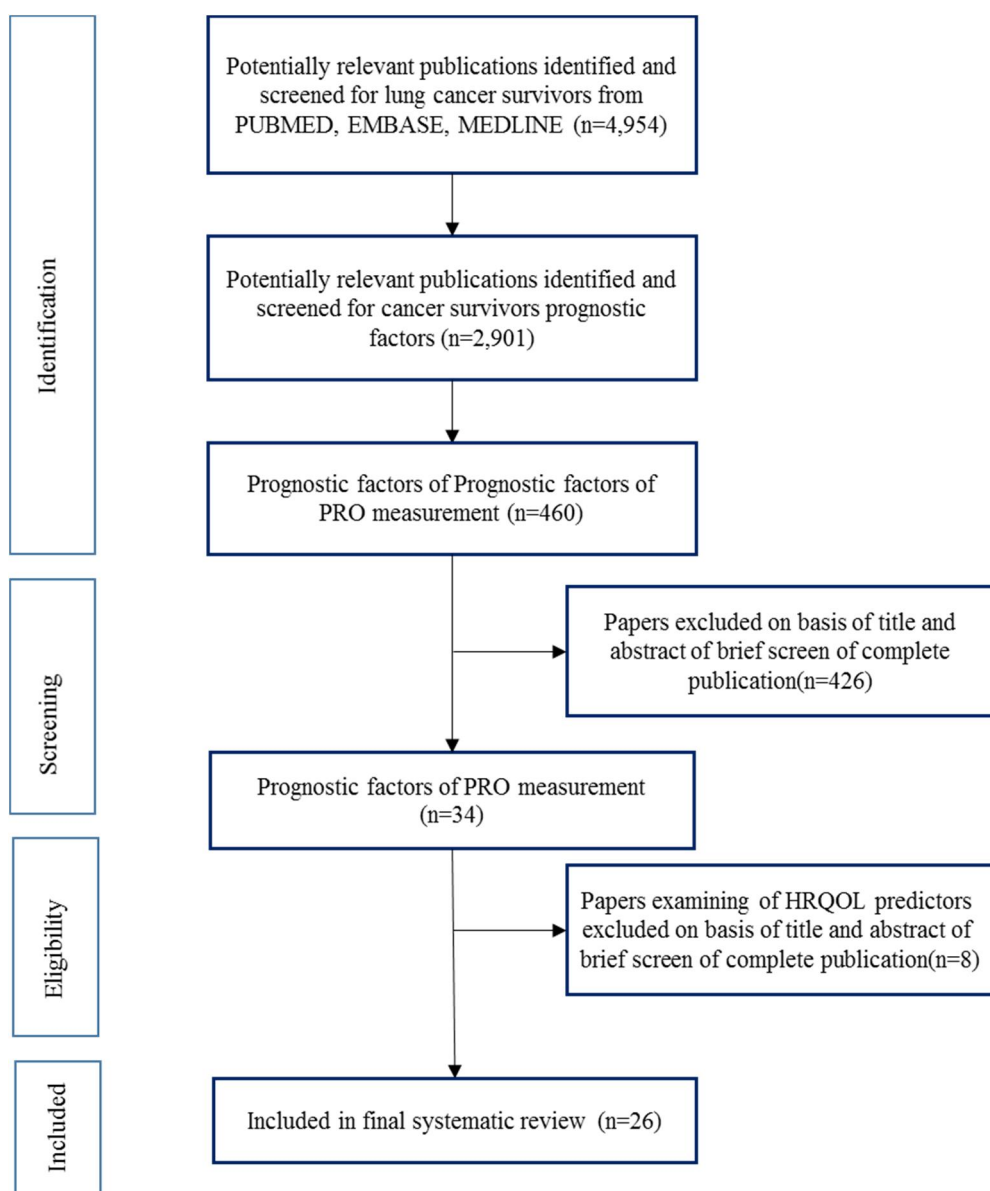


Figure III-1. Flow of selection of candidate prognostic factors from systematic review

## Approach 3

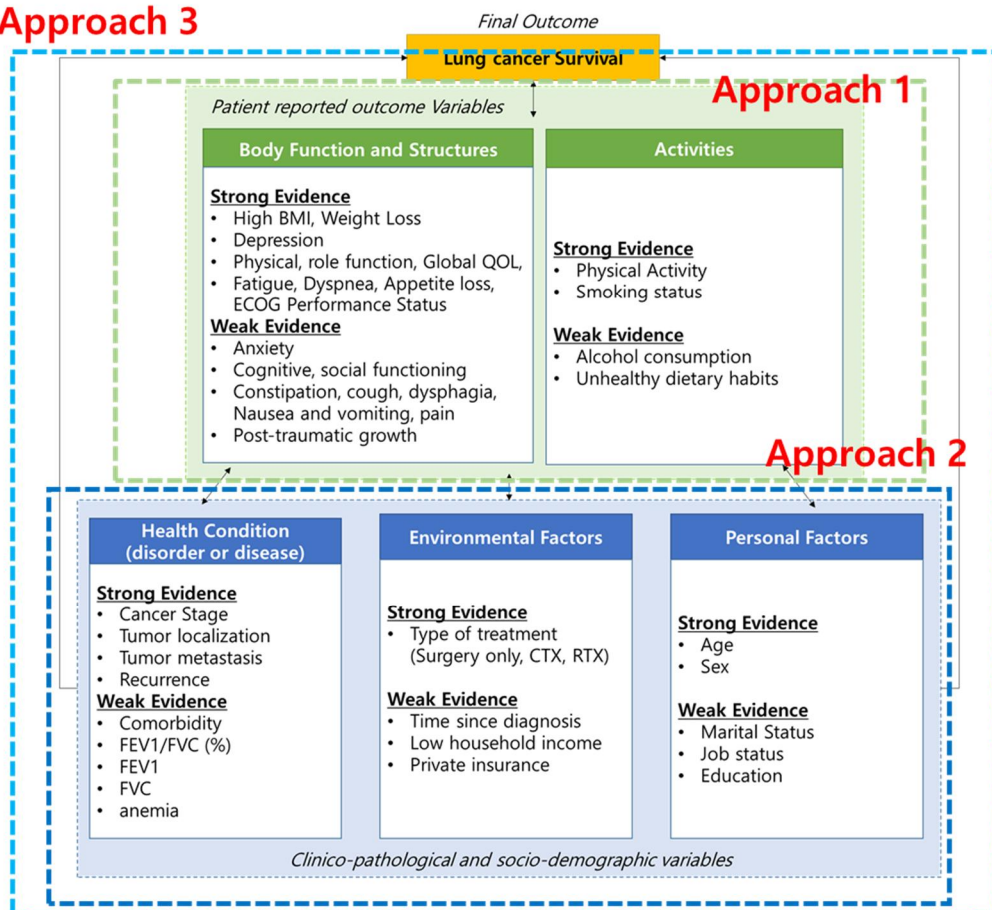


Figure III-2. Schematic diagram of candidate prognostic factors mapped with the bio-psychological framework based on the International Classification of Functioning, Disability and Health (ICF)

## **2. Model constructing feature sets with selecting prognostic factors**

From the summarized evidence, prognostic factors that are included in the evidence level of strong, weak, and inconsistent groups were considered as candidate predictive factors for predictive modeling of lung cancer survival. Among the selected prognostic factors, those that were questioned and checked in our study were finally chosen as candidate variables for modeling. For model constructing, we grouped the candidate prognostic factors into each model with three feature sets: 1) variables from “Health Condition”, “Environmental factors”, “Personal Factors” domains, 2) variables from “Body function and structures”, “Activities”, 3) feature sets including variables from 1) and 2) altogether (Box 1).

The selected variables from literature review were further included in the phased model according to the additional statistical analyses, in order to investigate whether the performance of the models improved significantly.

Box 1. Model constructing feature sets with candidate variables from literature review

- Feature set 1: variables from “Health Condition”, “Environmental factors”, “Personal Factors” domains
- Feature set 2: variables from “Body function and structure”, “Activities” domains
- Feature set 3: including variables from feature sets 1 with 2

## **B. Baseline characteristics**

### **1. Demographics of participants' characteristics and survival data**

Table III-1 summarizes the baseline demographic and clinical characteristics, as well as univariate analysis result with OS. The mean ages  $\pm$  SD were  $62.51 \pm 10.66$  and  $66.31 \pm 8.75$  in the non-event group and event group, respectively. Individuals more than 65 years experienced death more ( $p < 0.001$ ) and more females lived in comparison to males (58.4% versus 41.1%). Individuals with less than 3,000 USD monthly income also died more ( $p = 0.0014$ ). FEV1/FVC ratio  $< 0.7$  was also the candidate of death prognostic factor. Having local invasion of tumor, regional lymph node metastasis, recurrence, and the cancer stage also showed significant differences between the death and alive groups.

A total of 96 out of 809 cases (11.9%) died during the study. The person-years at risk data were accumulated for each patient from the date of the survey to the date of death. During the follow-up of 4509.2 person-years, we identified 96 deaths (11.9%) among the 809 subjects. In the 809 lung cancer survivors for whom there were available data, the median time from the diagnosis date to survey date was 6.0 ( $\pm 1.24$ ) years and the median survival time was 8.3 ( $\pm 2.01$ ) years.

Table III-1. Comparison of clinico-pathologic and socio-demographic characteristics between the event (dead) and no-event (alive) groups

Variable	No event		Event		p-value
	n	(%)	n=131	7.1%	
Age, years	62.51±10.66		66.31±8.75		<0.001
<65	393	55.12	33	34.38	0.0001
≥65	320	44.88	63	65.63	
Sex					
Female	177	94.65	10	5.35	0.0017
Male	536	86.17	86	13.83	
Monthly income (USD)					
≥ 3,000	207	94.10	13	5.90	0.0014
< 3,000	506	85.90	83	14.10	
Education					
≥ High school degree	185	90.24	20	9.76	0.2795
< High school degree	528	87.42	76	12.58	
Employment status					
Yes	285	90.48	30	9.52	0.0999
No	428	86.64	66	13.36	
Currently married					
Yes	655	88.04	89	11.95	0.7755
No	58	89.23	7	10.77	
FEV1/FVC	72.55±15.11		66.29±10.82		0.4710
(FEV1/FVC)*100 ≥ 0.7	431	92.1	37	7.9	<0.001
(FEV1/FVC)*100 < 0.7	244	81.1	57	18.9	
Local invasion of tumor					
No	253	92	22	8	0.017
Yes	460	86.3	73	13.7	
Regional lymph node metastasis					
No	508	89.8	58	10.2	0.042
Yes	205	84.7	37	15.3	
Stage					
Stage 0– I	464	65.08	46	47.92	0.0011
Stage II–III	249	34.92	50	52.08	
Recurrence					
No	630	92.51	551	7.49	<0.001
Yes	83	64.84	45	35.16	
Number of comorbidity					
0	318	87.60	45	12.40	0.6998
≥ 1	392	88.49	51	11.51	
Treatment type					
OP	431	88.7	55	11.3	0.217
OP+RT	39	81.3	9	18.8	
OP+CT	193	89.4	23	10.6	
OP+CT+RT	44	83	9	17	
Time since diagnosis	2.93 ±1.59		2.89±1.74		
≥ 3 years	306	89.21	37	10.79	0.4154
< 3 years	407	87.34	59	12.66	

## **2. Candidate selection from statistical analyses**

### **2.1.Univariate analysis of HRQOL mean scores between non-event and event groups**

Considering that we had to use the classification methods of MLTs, univariate of categorized HRQOL scores were also analyzed. The study result is shown in Table III-2. Appendix Table 17~19 summarizes the distribution of disease-free lung cancer survivors' HRQOL mean scores. EORTC QLQ-C30, physical functioning ( $p=0.001$ ), role functioning ( $p=0.001$ ), emotional functioning ( $p=0.008$ ), cognitive functioning, ( $p=0.015$ ), social functioning ( $p=0.004$ ), global QOL ( $p=0.018$ ), fatigue ( $p=0.002$ ), pain ( $p=0.032$ ), dyspnea ( $p<0.001$ ), appetite loss ( $p<0.001$ ), and financial difficulties ( $p<0.001$ ) showed significant differences between non-event and event groups.

In EORTC QLQ-LC 13 scales, lung cancer specific dyspnea ( $p<0.001$ ) and coughing ( $p<0.001$ ) were the only significant variables between the two groups, whereas sour mouth ( $p=0.09$ ) was marginally significant. Considering PTGI, which assesses the patients' post-traumatic growth, personal strength showed a significant difference between death and alive groups, while spiritual change and appreciation of life showed marginal significance. Finally, considering anxiety and depression, both the scales were significant ( $p=0.001$ ).



Table III-2. Comparison of EORTC QLQ-C30 HRQOL factors between the event (dead) and no-event (alive) groups

Variable		No event (n=713)		Event (n=96)		p-value
		n	(%)	n=131	7.1%	
Eortc-QLQ-C30						
Physical functioning	>33.33	696	88.9%	87	11.1%	p <0.001
	≤33.33	17	65.4%	9	34.6%	
Role functioning	>33.33	693	88.6%	89	11.4%	0.022
	≤33.33	20	74.1%	7	25.9%	
Emotional functioning	>33.33	700	88.4%	92	11.6%	0.133
	≤33.33	13	76.5%	4	23.5%	
Cognitive functioning	>33.33	699	88.1%	94	11.9%	0.937
	≤33.33	14	87.5%	2	12.5%	
Social functioning	>33.33	695	88.4%	91	11.6%	0.137
	≤33.33	18	78.3%	5	21.7%	
General health status	>33.33	681	88.4%	89	11.6%	0.229
	≤33.33	32	82.1%	7	17.9%	
Fatigue	<66.66	637	88.5%	83	11.5%	0.397
	≥66.66	76	85.4%	13	14.6%	
Nausea and vomiting	<66.66	696	88.1%	94	11.9%	0.855
	≥66.66	17	89.5%	2	10.5%	
Pain	<66.66	665	88.4%	87	11.6%	0.342
	≥66.66	48	84.2%	9	15.8%	
Dyspnea	<66.66	538	91.0%	53	9.0%	p <0.001
	≥66.66	175	80.3%	43	19.7%	
Insomnia	<66.66	610	88.5%	79	11.5%	0.399
	≥66.66	103	85.8%	17	14.2%	
Appetite loss	<66.66	641	89.4%	76	10.6%	0.002
	≥66.66	72	78.3%	20	21.7%	
Constipation	<66.66	659	88.3%	87	11.7%	0.536
	≥66.66	54	85.7%	9	14.3%	
Diarrhea	<66.66	685	88.6%	88	11.4%	0.049
	≥66.66	28	77.8%	8	22.2%	
Financial difficulties	<66.66	609	88.8%	77	11.2%	0.182
	≥66.66	104	84.6%	19	15.4%	

Table III-3. Comparison of EORTC QLQ-LC13 HRQOL factors between the event (dead) and no-event (alive) groups

Variable		No event (n=713)		Event (n=96)		p-value
		n	(%)	n=131	7.1%	
EORTC QLQ-LC13						
Dyspnea	<66.66	656	89.5%	77	10.5%	p <0.001
	≥66.66	57	75.0%	19	25.0%	
Coughing	<66.66	649	89.4%	77	10.6%	0.001
	≥66.66	64	77.1%	19	22.9%	
Hemoptysis	<66.66	708	88.2%	95	11.8%	0.715
	≥66.66	5	83.3%	1	16.7%	
Sore mouth	<66.66	692	88.5%	90	11.5%	0.073
	≥66.66	20	76.9%	6	23.1%	
Dysphagia	<66.66	693	88.5%	90	11.5%	0.072
	≥66.66	20	76.9%	6	23.1%	
Peripheral neuropathy	<66.66	641	88.4%	84	11.6%	0.469
	≥66.66	72	85.7%	12	14.3%	
Alopecia	<66.66	667	88.2%	89	11.8%	0.716
	≥66.66	45	86.5%	7	13.5%	
Pain in chest	<66.66	647	89.0%	80	11.0%	0.024
	≥66.66	66	80.5%	16	19.5%	
Pain in arm or shoulder	<66.66	617	88.3%	82	11.7%	0.764
	≥66.66	96	87.3%	14	12.7%	
Pain in other parts	<66.66	643	88.6%	83	11.4%	0.394
	≥66.66	70	85.4%	12	14.6%	

Table III-4. Comparison of PTGI and HADS factors between the event (dead) and no-event (alive) groups

		No event (n=713)		Event (n=96)		p-value
Variable		n	(%)	n=131	7.1%	
PTGI						
Relation to others (35)	≥23	297	89.5%	35	10.5%	0.331
	<23	416	87.2%	61	12.8%	
New possibilities (25)	≥18	164	92.7%	13	7.3%	0.032
	<18	501	86.7%	77	13.3%	
Personal strength (20)	≥15	223	94.1%	14	5.9%	0.001
	<15	490	85.7%	82	14.3%	
Spiritual change (10)	≥5	367	90.2%	40	9.8%	0.071
	<5	346	86.1%	56	13.9%	
Appreciation of life (15)	≥11	328	90.9%	33	9.1%	0.031
	<11	385	85.9%	63	14.1%	
HADS						
Anxiety	<8	575	90.7%	59	9.3%	p <0.001
	≥8	134	78.8%	36	21.2%	
Depression	<8	445	90.8%	45	9.2%	0.002
	≥8	262	83.7%	51	16.3%	

## **2.2. Univariate analysis of BMI, weight change, and MET of lung cancer survivors**

Table III-5 summarizes the distribution of disease-free lung cancer survivors' BMI, weight change, and PA measured with MET. PA and BMI were measured at a median of 29.8 months after diagnosis. In comparison to the reference category, subjects with a BMI < 23.5 kg/m<sup>2</sup> had significantly higher proportions of death and those who gained ≥ 4 kg after diagnosis had a higher risk of death in comparison to the subjects who gained < 4 kg. For disease-free lung cancer survivors who engaged in < 12.5 versus ≥ 12.5 MET-hours per week of PA showed significant difference between non-event and event groups. Other lifestyle factors did not show any significant difference between the alive and death groups.

Table III-5. Comparison of lifestyle factors between the event (dead) and no-event (alive) groups

Variable	No event		Event		p-value
	n=1,719	92.9%	n=131	7.1%	
Present BMI(kg/m2)					
< 23	303	42.50	51	53.13	0.0488
≥ 23	410	57.50	45	46.88	
BMI (kg/m2) before operation					
< 23	299	41.99	56	58.95	0.0018
≥ 23	413	58.01	39	41.05	
BMI (kg/m2) changed (before operation-present)					
Reduction	247	34.64	28	29.17	0.0024
Maintenance	306	42.92	31	32.29	
Increase	160	22.44	37	38.54	
Alcohol now					
No	548	76.86	75	78.13	0.78
Yes	165	23.14	21	21.88	
Alcohol experience					
No	271	38.01	29	30.21	0.1374
Yes	442	61.99	67	69.79	
Present smoking status					
No	661	92.71	89	92.71	0.9996
Yes	52	7.29	7	7.29	
MET					
< 12.5	395	89.98	44	10.02	0.075
≥ 12.5	317	85.91	52	14.09	

### **3. Final candidate variable selection for phased modeling**

Based on the literature review and statistical analyses, the final candidate variable selections for three phased modeling feature sets were constructed. According to those of three modeling feature sets, we further developed each of model based on the cox regression model and four MLTs. Final candidate variable selected from both literature review, and statistical analyses are shown in Table III-6. Among 121 available variables in the data set, which comprised of health condition, environmental factors, personal factors, body function and structures, 4 health condition, 6 clinically relevant variables, 19 HRQOL variables, and 2 lifestyle factors were preliminarily selected.

In our model development process, though the ‘FEV1/FVC ratio’ and ‘weight change’ were the significant prognostic factors and also suggested as strong evidence values, in addition to the effects of other prominent covariates or multi-collinearities with HRQOL dyspnea function and BMI, we did not include two variables for modeling. In addition, because the recurrence factor was also regarded as outcome variable, we did not put the variable in the modeling process.

Table III-6. Final candidate variables from both literature review and statistical analyses

Factors	Variables	Literature review evidence level	Significant variables form statistical analyses
Health Condition	Cancer stage	Strong	O
	Local invasion of tumor	Strong	O
	Regional Lymph node metastasis	Strong	O
	Number of comorbidity	Strong	NS
	Recurrence	Strong	O
Environmental Factors	Time since diagnosis	Weak	NS
	Type of treatment	Strong	NS
	Private insurance	Weak	NA
	Low household income	Weak	O
Personal Factors	Age	Strong	O
	Sex or Gender	Strong	O
	Job status	Weak	Marginally significant
	Education	Weak	NS
	Marital status	Weak	NS
Body function and structures	BMI(kg/m2) before operation	Strong	O
	weight change	Weak	O
	Anxiety	Strong	O
	Depression	Strong	O
	Physical functioning	Strong	O
	Role functioning	Strong	O
	Emotional functioning	Weak	NS
	Cognitive functioning	Weak	NS
	Social functioning	Weak	NS
	General health QOL	Strong	NS
	Fatigue	Strong	NS
	Nausea and vomiting	Weak	NS
	Pain	Weak	NS
	Dyspnea	Strong	O
	Appetite loss	Strong	O
	Diarrhea	Weak	O
	Constipation	Weak	NS
	Coughing	Weak	O
	Peripheral neuropathy	Weak	NS
	Post-traumatic growth	Weak	O
	ECOG Performance	Strong	NA
Activities	Physical activity	Strong	O
	smoking status	Weak	NS
	Alcohol consumption	Weak	NS
	Unhealthy dietary habits	Weak	NS

**Abbreviation:** NS, Non-significant; NA, Not applicable; O, significant

## 4. Result of data preprocessing

### 4.1. Missing imputation

Prior to imputing the missing values, we first investigate the missing number of each variable. After identifying the missing values, using the code of the library (DMwR) and function of “KnnImputation” (data, k = 5), missing values were replaced and imputed. The “before” and “after” missing values plotting are shown in the below figure. The red points are the proportion of missing values, and we can observe that after KNN imputation, there were no more missing values in this data. (Figure III-2)

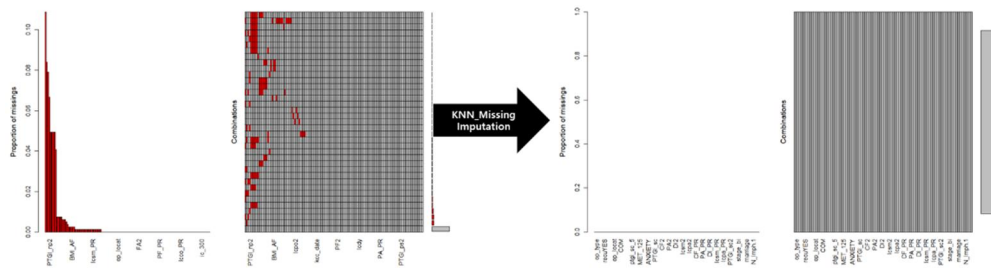


Figure III-2. Missing imputation process



#### 4.2. Data SMOTE and hold-out sampling

After missing imputations were completed, SMOTE to balance the whole data were also conducted. The data set has an event group of 576 and 480 live patients. Based on the SMOTE data, we also split the data into 70% as a training set and 30% as a validation set. The final result of SMOTE and data splitting is shown in Table III-7. No statistically significant difference between the training data set (n=739) and test data set (n=317) were found (Appendix).

Table III-7. Comparison of original data structure of alive and death groups with SMOTE data

	Original data N=819	SMOTE	
		Training set N=739 (70%)	Test set N=317(30%)
Case (alive)	713 (88.1)	406 (70.5)	170 (29.5)
Event (death)	96 (11.9)	333 (69.4)	147 (30.6)

## C. Model development

Phased modeling was based on the below mentioned variables, based on the result of literature review and univariate analyses, performed as a step by step phased modeling. We determined to use the HRQOL prognostic factors for model development as the predictors in the final model and attempted to identify the best fitting model which improves the model performance in comparison to traditional clinical variables included in the model. Finally, the three models developed, candidate variables selected from literature review, and univariate analyses for each model are as follows:

Box 2. Candidate variables selected for three types of feature sets

- **Feature set 1:** age, sex, monthly income, stage, regional lymph node metastasis, time since diagnosis
- **Feature set 2:** physical function, role function, appetite loss, dyspnea, diarrhea, lung cancer specific cough, pain in chest, new possibility, personal strength, appreciation of life, anxiety, depression, before BMI, MET, time since diagnosis
- **Feature set 3:** age, sex, monthly income, stage, regional lymph node metastasis, time since diagnosis, physical function, role function, appetite loss, dyspnea, diarrhea, lung cancer specific cough, pain in chest, new possibility, personal strength, appreciation of life, anxiety, depression, before BMI, MET, time since diagnosis

## **1 . Cox model development**

Multivariate analyses of lung cancer survival prediction model Cox-1, Cox-2, and Cox-3 were performed using the Cox regression model (Table 2). In the multivariate analysis, the original combination of prognostic factors in models Cox-1, Cox-2, and Cox-3 showed a good or better than other combinations of prognostic factors selected by stepwise variable selection methods. The final model Cox-1 includes age, sex, stage of lung cancer, income, treatment type, and regional lymph node metastasis. The application of prognostic scores of individual survivors allows for the calculation of expected lung cancer survival.

Those categorized as predictors were then used as variables for each model Cox-1, 2, and 3 as shown in Table III-8. The best fitted model with the least AIC was model Cox-3, containing factors from model Cox-1 and 3. The AIC of model 1-A was 5021.168 and that of model Cox-2 was 4922.56. The AIC of model 3 was the lowest at 4869.012, which indicates a better fit. The C-statistic showed a similar pattern, where the scores of C-statistic was higher for model 3, and the differences were statistically significant between model Cox-1 and Cox-2, thus model Cox-3 could be suggested as the best performing model.

Table III-8. Possible models in phased cox modeling for lung cancer survivors

	Variables used in model	AIC	C-statistics
Model Cox-1	Age over 65 years, Male (versus female), stage II–III (versus stage 0–I), income < 3,000 (versus ≥ 3,000), regional lymph node metastasis (versus no),	5021.168	0.699
Model Cox-A	time since diagnosis ≥ 3 years (versus < 3 years), BMI (kg/m <sup>2</sup> ) before operation ≥ 23 (versus < 23), problematic role functioning ≤ 33.33 (versus > 33.33), problematic dyspnea ≥ 66.66 (versus < 66.66), personal strength < 15/20 (versus ≥ 15), appreciation of life < 18 (versus ≥ 18)	4922.56	0.767
Model Cox-A	time since diagnosis ≥ 3 years (versus < 3 years), BMI (kg/m <sup>2</sup> ) before operation ≥ 23 (versus < 23), problematic role functioning ≤ 33.33 (versus > 33.33), problematic dyspnea ≥ 66.66 (versus < 66.66), personal strength < 15/20 (versus ≥ 15), appreciation of life < 18 (versus ≥ 18), Male (versus female), stage II–III (versus stage 0–I), treatment type	4869.012	0.809

### 1.1. Prediction model based on Cox regression analysis

As concluded in the previous section, each model Cox-1, 2, and 3 containing all candidate predictor variables were constructed. Table III-9 through Table III-12 show the multivariate adjusted odds ratios for each of the best fitting model. With adjustment for the independent indicators of survival, final multiple proportional hazard regression analyses of model Cox-1 shows that individuals over 65 years (aHR, 1.32; 95% CI, 1.08–1.67), female (aHR, 0.53; 95% CI, 0.37–0.72), stage II–III (aHR, 1.36; 95% CI, 1.03–1.78), income < 3,000 (aHR, 1.36; 95% CI, 1.03–1.78), and regional lymph node metastasis (aHR, 1.31; 95% CI, 1.03–1.67) did not lose the independent prognostic power of survival (Table III-9).

In model Cox-2, time since diagnosis  $\geq 3$  years (aHR, 0.82; 95% CI, 0.67–0.99), BMI (kg/m<sup>2</sup>) before operation  $\geq 23$  (aHR, 0.54; 95% CI, 0.44–0.67), problematic role functioning (aHR, 2.37; 95% CI, 1.76–3.19), problematic dyspnea (aHR, 1.57; 95% CI, 1.27–1.93), personal strength (aHR, 2.52; 95% CI, 1.69–3.75), and appreciation of life < 18 (aHR, 1.63; 95% CI, 1.25–2.12) showed the best prognostic factors for lung cancer survival (Table III-10).

In model Cox-3, BMI (kg/m<sup>2</sup>) before operation  $\geq 23$  (aHR, 0.54; 95% CI, 0.43–0.67), problematic role functioning (aHR, 2.20; 95% CI, 1.63–2.97), problematic dyspnea (aHR, 1.47; 95% CI, 1.19–1.81), personal strength (aHR, 2.23; 95% CI, 1.50–3.32), appreciation of life < 18 (aHR, 1.603; 95% CI, 1.23–2.09), stage II–III (aHR, 1.32; 95% CI, 1.06–1.63), and sex (aHR, 0.47; 95% CI, 0.35–0.65) were the final predictive values for lung cancer survival (Table III-11).

Table III-9. Lung cancer survivors' mortality prediction model Cox-1

Prognostic factor	$\beta^a$	aHR	95% CI	P -value
Age over 65 years (Ref. < 65)	0.280	1.324	1.084-1.671	0.006
Female (Ref. male)	-0.639	0.528	0.386-0.722	<.0001
Stage II–III (Ref. stage 0– I )	0.306	1.358	1.092-1.688	0.006
Income < 3,000 (Ref. $\geq$ 3,000)	0.305	1.357	1.032-1.784	0.0288
Regional lymph node metastasis (Ref. no)	0.271	1.312	1.029-1.672	0.0283
First Entered Model AIC	5071.65			
Best Optimized Model AIC <sup>b</sup>	5021.168 <sup>b</sup>			

**Abbreviations:** HR, Hazard Ratio; CI, Confidential Interval; OP, Operation; RT, Radiotherapy; CT, Chemotherapy

**a. Mortality prediction model score:**  $0.280 \times (\text{Age over 65 years}) - 0.639 \times (\text{Female}) + 0.306 \times (\text{Stage II–III}) - 0.24 \times (\text{OP+RT Treatment}) - 0.375 \times (\text{OP+CT Treatment}) + 0.383$

**b.** Lower value indicates better fit.

Table III-10. Lung cancer survival prediction model Cox-2

Prognostic factor	$\beta^a$	aHR	95% CI	P -value
Time since diagnosis $\geq 3$ years (Ref. < 3 years)	-0.204	0.815	0.667- 0.997	0.0467
BMI (kg/m <sup>2</sup> ) Before operation $\geq 23$ (Ref. < 23)	-0.613	0.542	0.436- 0.673	<.0001
Problematic physical functioning $\leq 33.33$ (Ref. > 33.33)	0.862	2.367	1.758- 3.186	<.0001
Problematic dyspnea $\geq 66.66$ (Ref. < 66.66)	0.449	1.567	1.274- 1.927	<.0001
Personal Strength < 15/20 (Ref. $\geq 15$ )	0.924	2.52	1.694- 3.748	<.0001
Appreciation of life < 18 (Ref. $\geq 18$ )	0.488	1.63	1.251- 2.123	0.0003
First Entered Model AIC	5051.301			
Best Optimized Model AIC <sup>b</sup>	4922.56			

**Abbreviations:** aHR, adjusted Hazard Ratio; CI, Confidential Interval; OP, Ooperation; RT, Radiotherapy; CT, Chemotherapy

- a. Mortality Prediction model score:  $-0.204 \times (\text{Time since diagnosis} \geq 3 \text{ years}) - 0.613 \times (\text{BMI (kg/m}^2\text{) before operation} \geq 23) + 0.862 \times (\text{Problematic role function}) - 0.449 \times (\text{Problematic dyspnea}) + 0.924 \times (\text{Personal strength} < 15) + 0.488 \times (\text{Appreciation of life} < 18)$
- b. Lower value indicates better fit.

Table III-11. Lung cancer survival prediction model Cox-3

Prognostic factor	$\beta$	aHR	95% CI	P -value
BMI (kg/m <sup>2</sup> ) before operation $\geq 23$ (Ref. < 23)	-0.616	0.540	0.434-0.672	<.0001
Problematic role functioning $\leq 33.33$ (Ref. > 33.33)	0.788	2.200	1.627-2.974	<.0001
Problematic dyspnea $\geq 66.66$ (Ref. < 66.66)	0.384	1.469	1.192-1.810	0.0003
Personal Strength < 15/20 (Ref. $\geq 15$ )	0.802	2.230	1.500-3.317	<.0001
Appreciation of life < 18 (Ref. $\geq 18$ )	0.472	1.603	1.230-2.088	0.0005
Stage II–III (Ref. stage 0–I)	0.275	1.316	1.064-1.627	0.0112
Female (Ref. male)	-0.746	0.474	0.346-0.650	<.0001
First Entered Model AIC	5051.301			
Best Optimized Model AIC <sup>b</sup>	4869.012			

**Abbreviation:** aHR, adjusted Hazard Ratio; CI, Confidential Interval;

- a. **Mortality Prediction model score:**  $-0.616 \times (\text{BMI (kg/m}^2\text{) before operation} \geq 23) + 0.788 \times (\text{Problematic role functioning}) - 0.384 \times (\text{Problematic dyspnea}) + 0.802 \times (\text{Personal strength} < 15) + 0.472 \times (\text{Appreciation of life} < 18) - 0.746 \times (\text{Female}) + 0.275 \times (\text{Stage II–III})$
- b. Lower value indicates better fit.



## 1.2. Final prediction model equation for Cox models

The final prediction model chosen as Cox regression model can be elaborated by the following equations:

Box 3. Final prediction model equation for Cox Models

<p><b>Model Cox-1</b> prediction score = <math>0.280 \times (\text{Age over 65 years (yes[1], no[0])}) - 0.639 \times (\text{Female (yes[1], no[0])}) + 0.306 \times (\text{Stage II–III (yes[1], no[0])}) - 0.24 \times (\text{OP+RT Treatment (yes[1], no[0])}) - 0.375 + 0.271 \times (\text{Regional lymph node metastasis (yes[1], no[0])})</math></p>
<p><b>Model Cox-2</b> prediction score = <math>-0.204 \times (\text{Time since diagnosis} \geq 3 \text{ years (yes[1], no[0])}) - 0.613 \times (\text{BMI (kg/m}^2\text{) before operation} \geq 23 \text{ (yes[1], no[0])}) + 0.862 \times (\text{Problematic role functioning (yes[1], no[0])}) - 0.449 \times (\text{Problematic dyspnea (yes[1], no[0])}) + 0.924 \times (\text{Personal strength} &lt; 15 \text{ (yes[1], no[0])}) + 0.488 \times (\text{Appreciation of life} &lt; 18 \text{ (yes[1], no[0])})</math></p>
<p><b>Model Cox-3</b> prediction score = <math>-0.616 \times (\text{BMI (kg/m}^2\text{) before operation} \geq 23 \text{ (yes[1], no[0])}) + 0.788 \times (\text{Problematic role functioning (yes[1], no[0])}) - 0.384 \times (\text{Problematic dyspnea (yes[1], no[0])}) + 0.802 \times (\text{Personal strength} &lt; 15 \text{ (yes[1], no[0])}) + 0.472 \times (\text{Appreciation of life} &lt; 18 \text{ (yes[1], no[0])}) - 0.746 \times (\text{Female (yes[1], no[0])}) + 0.275 \times (\text{Stage II–III (yes[1], no[0])})</math></p>

## **2. Decision tree model development**

### **2.1. Assessment of the relative importance and model developing**

Similar to stepwise variable selection in Cox proportional hazard regression model analysis, DT methods for each of the three models also conducted variable selection of the most relevant input variables, which were used to formulate each prognostic factor model. The normalized mutual information index (NMI), which is based on mutual importance, in addition to its role in calculating the correlation coefficient, is also used to determine the importance of an explanatory prognostic factor for the prediction of overall survival [81, 82]. Mutual information is a quantitative measure for the mutual variables' dependences.

The significance of the selected prognostic factors for each of the three models based on DT algorithms were compared. In model DT-1, treatment type and regional lymph node metastasis were identified as the most important factors in the prediction of overall survival, while sex, stage, income, and age over 65 years were the next important variables. Time since diagnosis more than 3 years was the least important variable. In model DT-2, problematic dyspnea, was the most important variable, while personal strength, depressive symptoms, new possibility score from PTGI, BMI before operation, problematic role and physical functioning, chest pain symptoms, and anxiety were the next important variables. In model DT-

3, most of the highest variables were HRQOL variables, where problematic dyspnea was also the highest variable (44.77%).

Table III-12. Importance of prognostic factors by normalized mutual information index of DT models

Variables	Normalized mutual information index (%)		
	Model DT-1	Model DT-2	Model DT-3
Regional lymph node metastasis	27.64	-	NS
Sex	10.80	-	10.33
Stage II –III	26.16	-	6.24
Income < 3,000	6.09	-	NS
Age over 65 years	3.60	-	4.66
Time since diagnosis $\geq 3$ years	6.67	NS	4.66
Problematic dyspnea $\geq 66.66$	-	32.75	32.75
Lower personal strength	-	21.10	23.59
Depression	-	4.15	6.83
Lower new possibility	-	NS	NS
BMI (kg/m <sup>2</sup> ) before operation $\geq 23$	-	1.29	6.55
Problematic role functioning $\leq 33.33$	-	8.44	6.37
Problematic physical functioning $\leq 33.33$	-	NS	NS
Problematic Appetite loss $\geq 66.66$	-	7.66	3.19
Problematic chest pain $\geq 66.66$	-	6.53	7.28
Anxiety	-	3.56	3.16
MET	-	1.01	4.03
Appreciation of life < 18	-	17.34	7.00

## **2.2. Selecting CP value for decision tree pruning using “rpart” packages**

When deriving a DT model, all observations in lung cancer training set start from the root node. Further, for each of the prognostic factors, the optimal binary split is determined. In node impurity-based DT models, optimality is defined as the split resulting in the largest decrease in node impurity. To identify the number of DT pruning, cross-validation based on the training set were conducted for each of the three models, selecting complexity parameter (CP) value by choosing the lowest level of the minimum “xerror value” using splitting rules [73, 83].

Even though there were diverse splitting packages, such as “tree,” “rpart,” and “party,” we used rpart packages’ plotcp function to plot the CP table and rpart tree fitting on each of the three models based on the training data. The results of CP tables and plots of models DT-1, DT-2, and DT-3 are shown in Appendix Figure 2-4. In order to identify the minimum xerror values for each model, we used the function of minsplit, which showed the minimal number of observations. For model DT-1, the best splitting parameter was five times, while the ideal tree size was five splitting for model DT-2. In model DT-3, 24 times splitting was identified as the most appropriate number to prune the best fitting tree. After pruning, the DT algorithms for models DT-1 through 3 were plotted. The plotted DT algorithms are shown in Figure III-3. The node that splits in the DT model can provide an indication of what specific levels of the prognostic factors were statistically associated with lung cancer mortality.

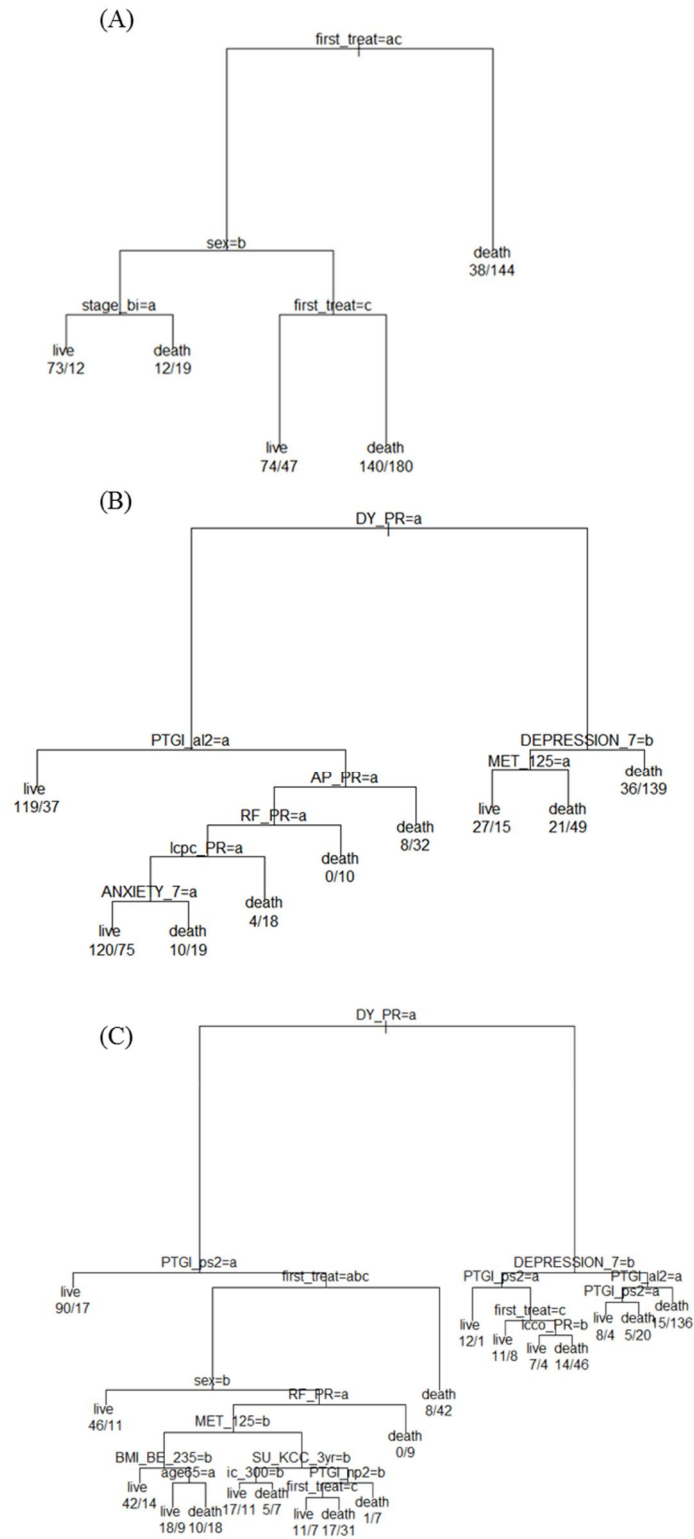


Figure III-3. Plots of decision tree models  
(A) Model DT-1 (B) Model DT-2 (C) Model DT-3

### **3. Random forest model development**

To predict lung cancer mortality based on the RF algorithm, variable importance scores for RF were computed by assessing the increase in prediction error, when the values of prognostic factors predicting mortality are replaced across the OOB data [78, 84]. The score of variable importance was calculated for each constructed tree, averaged across the entire ensemble, and divided by the SD. The plots investigating variable importance predicting survival were calculated using mean decrease accuracy and mean decrease Gini scores. The results are shown in Figure III-4.

In this study, the RF comprised of 100 fully grown trees from the training set. Prior to constructing the prediction model, optimal mtry should be identified to avoid over-fitting. Starting with the default value of mtry, the optimal value (with OOB error estimate) of mtry from package of randomForest should be identified. For each of the bootstrap samples to avoid over-fitting, the best split among all the predictors which were randomly explored by “mtry” meaning at least error rate and avoid overfitting for model. As the number of trees increase, the error rate decreases and the plot of error rate changes based on the increasing number of trees. Further, a final model prediction was undertaken from the training set for each sample tree of each of the three models.

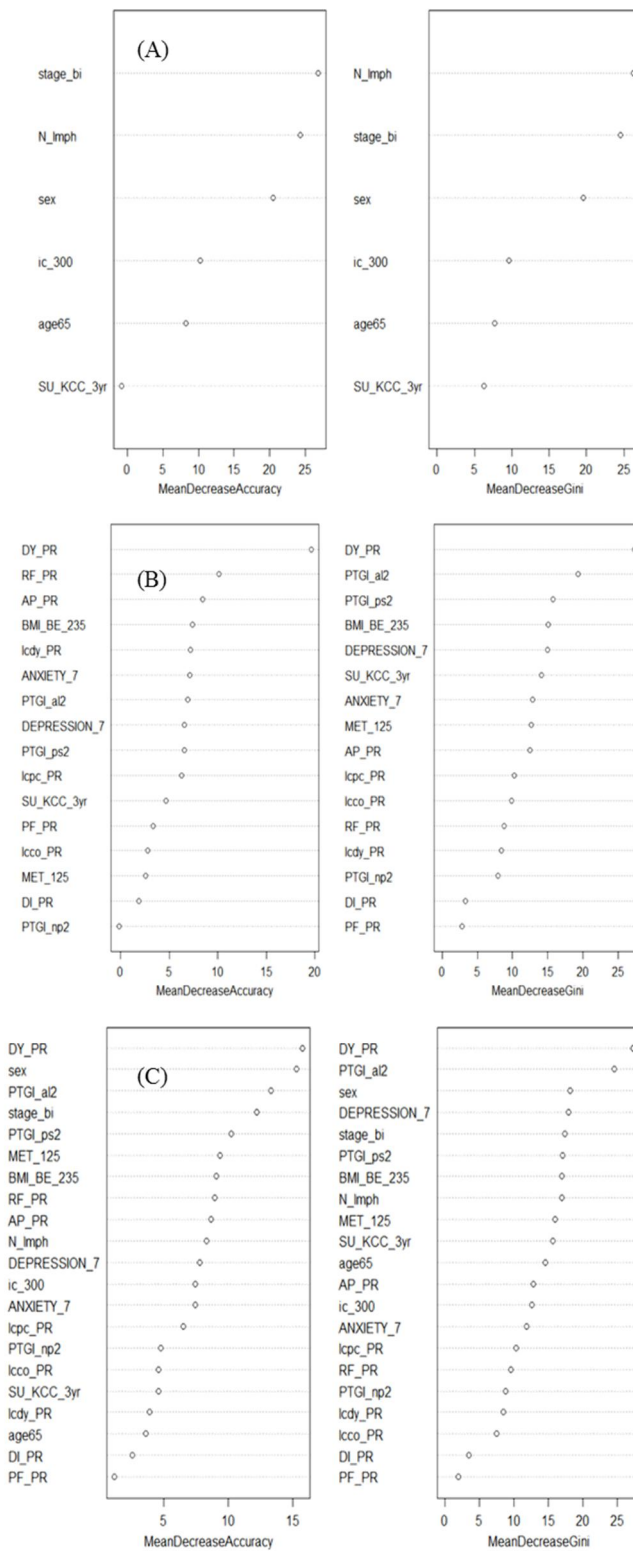


Figure III-4. Random forest variable importance plots  
(A) Model RF-1 (B) Model RF-2 (C) Model RF-3

#### 4. Bagged decision tree model development

For the bagged decision tree development, we created 78 bootstrapping samples showing the lowest OOB error rate, where the multiple model prediction was trained with the training set. In the bagged model, optimal predictors using the bootstrap bagging algorithms were first selected. Further, based on the model development, including selected prognostic factors, a combination of results of each model were used with test set to predict. Variable importance (%) was also analyzed as an NMI. The error rate of OOB according to the number of bootstrapping samples was also investigated and the results are shown below.

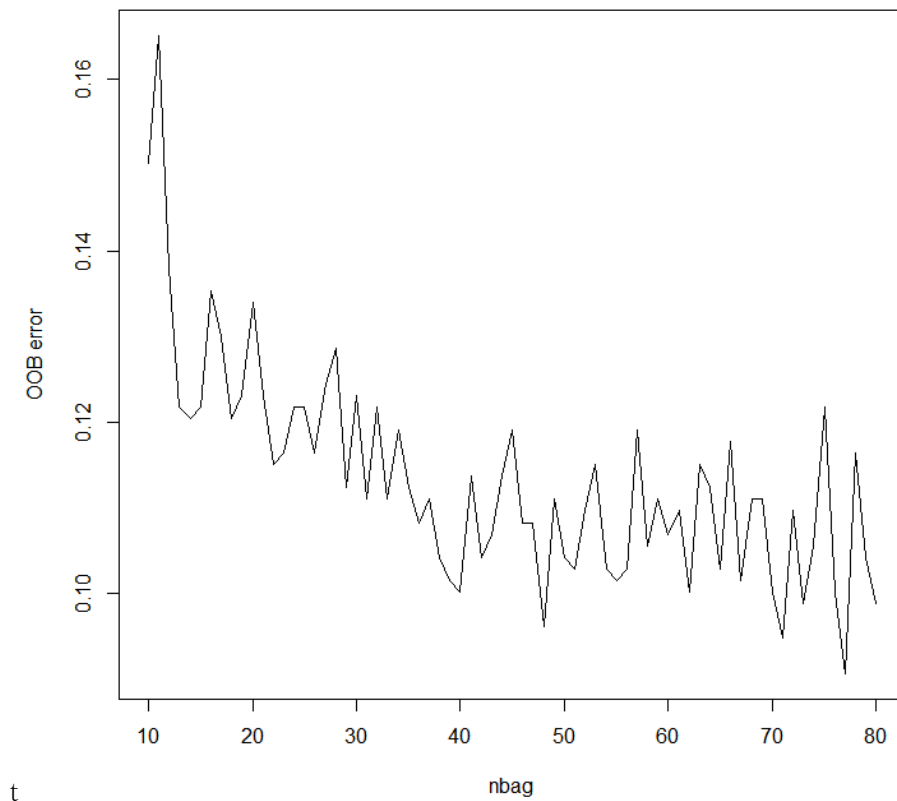


Figure III-5. Out-of-bag (OOB) error rate according to number of bootstraps



## 5. AdaBoost model development

First, model training with the training set was constructed for each of the three models (model AB-1, AB-2, and AB-3). The number of trees (n.trees) suitable for the data is selected with training data set. Function of R to tune the suitable number of tree is as follows: `tune.boost <- errorevol (boost, newdata = train)`. To develop the AdaBoost model, the default 100 data the input data. Repeating bootstrapping several times results in a stronger classifier construction.

Treatment type (NMI: 28.79%), regional lymph node metastasis (NMI: 13.37%), sex (NMI: 10.92%), stage (NMI: 14.22 %), income (9.00%), age over 65 years (NMI: 8.45%), and time since diagnosis (NMI: 6.42%) affected model AB-1 for the prediction of lung cancer mortality. For model AB-2, dyspnea (NMI: 11.94%), depression (NMI: 10.20%), diarrhea (NMI: 4.53%), appetite loss (NMI: 5.27%), chest pain (NMI: 4.51%), cough (NMI: 3.87%), and anxiety (NMI: 4.07%) affected lung cancer mortality. For model AB-3, treatment type (NMI: 12.39%), income (NMI: 5.90), age (NMI: 6.04), dyspnea (NMI: 8.01%), appetite loss (NMI: 3.04%), and anxiety (NMI: 3.58%) affected lung cancer mortality. (Table III-13)

Table III-13. Importance of prognostic factors by normalized mutual information index

Variable	Normalized mutual information index (%)		
	Model AB-1	Model AB-2	Model AB-3
Treatment type	28.79	-	12.39
Regional lymph node metastasis	13.37	-	NS
Sex	10.92	-	NS
Stage II –III	14.22	-	NS
Income	9.00	-	5.90
Age over 65 years	8.45	-	6.04
Time since diagnosis $\geq 3$ years	6.42	-	NS
Problematic dyspnea	-	11.94	8.01
Lower personal strength	-	NS	NS
Depression $\geq 8$	-	10.20	NS
Lower new possibility	-	NS	NS
BMI (kg/m <sup>2</sup> ) before operation	-	NS	NS
Problematic role functioning	-	NS	NS
Problematic physical functioning	-	NS	NS
Problematic diarrhea	-	4.53	0.75
Problematic appetite loss	-	5.27	3.04
Problematic chest pain	-	4.51	NS
Problematic coughing	-	3.87	NS
Anxiety	-	4.07	3.58
MET	-	NS	NS
Appreciation of life	-	NS	NS

## **6. Developed models applied with MLTs**

The major variables affecting the survival of survivors of lung cancer are summarized according to the classification of prognostic factors in each the study model. (Table III-14~18) In model Cox-1, five variables were identified as important influential prognostic factor except for 'time since diagnosis' variable. (Table III-14) In cox-2 models based feature set 2, variables such as BMI, 'role functioning', 'dyspnea', 'personal strength' and 'appreciation of life' were important variables. When we compared the number of variables used in the development of Cox predictive models (Cox-1= 5, Cox-2=5, Cox-3=8).

In model DT-1, with the 6 variables can be the likelihood of the most optimal model to predict a survival group. In model DT-2, BMI, anxiety, depression, role function, dyspnea, appetite loss, personal strength, MET were investigated as the important variables. In DT model based feature set 3, time since diagnosis, sex, dyspnea, personal strength, and MET used in the development of predictive models (Table III-15).

In model RF-1, the same 6 variables which were selected from DT were the important variables. In model RF-2, BMI, anxiety, depression, role function, dyspnea, appetite loss, personal strength, new possibilities, MET were investigated as the important variables. In RF model based feature set 3, cancer stage, income, sex, dyspnea, anxiety, personal strength, and MET used in the development of predictive models (Table III-16).

In model Bag-1, also the same 6variables which were selected from DT and RF were the important variables. In model Bag-2, BMI, anxiety, depression,

role function, dyspnea, appetite loss, personal strength, new possibilities, MET were investigated as the important variables. In bag model based feature set 3, cancer stage, age, income, sex, dyspnea, anxiety, personal strength, appreciation of life, new possibilities, and MET used in the development of predictive models (Table III-17).

In model AdaBoost-1, also the same 6variables which were selected from DT and RF were the important variables. In model AdaBoost-2, BMI, anxiety, depression, dyspnea, appetite loss, lung cancer cough, MET were investigated as the important variables. In AdaBoost model based feature set 3, age, BMI, dyspnea, anxiety, appetite loss were used in the development of predictive models (Table III-18).

Table III-14. Selected important variables based on Cox models

Factors	Variables	Model Cox-1	Model Cox-2	Model Cox-3
Health Condition	Cancer stage	O		O
	Regional lymph node metastasis	O		
Environmental Factors	Time since diagnosis			O
	Low household income	O		
Personal Factors	Age	O		
	Sex	O		O
Body function and structures	BMI(kg/m <sup>2</sup> ) before operation		O	O
	Role functioning		O	O
	Dyspnea		O	O
	Personal strength		O	O
	Appreciation of life		O	O

Table III-15. Selected important variables based on DT models

Factors	Variables	Model DT-1	Model DT-2	Model DT-3
Health Condition	Cancer stage	O		
Environmental Factors	Time since diagnosis	O		O
	Low household income			
Personal Factors	Sex	O		O
	BMI(kg/m2) before operation	O	O	
	Anxiety		O	
	Depression		O	
	Physical functioning			
	Role functioning	O	O	
Body function and structures	Dyspnea	O	O	O
	Appetite loss		O	
	Diarrhea			
	Lung cancer specific cough			
	Pain in chest			
	New possibility			
	Personal strength	O	O	O
	Appreciation of life	O	O	
Activities	Physical activity (MET)		O	O

Table III-16. Selected important variables based on random forest models

Factors	Variables	Model RF- 1	Model RF- 2	Model RF- 3
Health Condition	Cancer stage	O		O
	Regional lymph node metastasis	O		O
Environmental Factors	Time since diagnosis	O		
	Low household income	O		O
Personal Factors	Age	O		
	Sex	O		O
Body function and structures	BMI(kg/m2) before operation		O	O
	Anxiety		O	O
	Depression		O	
	Physical functioning			
	Role functioning		O	
	Dyspnea		O	O
	Appetite loss		O	
	New possibility			
	Personal strength		O	O
	Appreciation of life		O	
Activities	Physical activity (MET)		O	O

Table III-17. Selected important variables based on Bagging models

Factors	Variables	Model Bag-1	Model Bag-2	Model Bag-3
Health Condition	Cancer stage	O		O
	Regional lymph node metastasis	O		O
Environmental Factors	Time since diagnosis	O		O
	Low household income	O		O
Personal Factors	Age	O		O
	Sex	O		O
Body function and structures	BMI(kg/m2) before operation		O	O
	Anxiety		O	O
	Depression		O	O
	Physical functioning			
	Role functioning		O	O
	Dyspnea		O	O
	Appetite loss		O	
	Diarrhea			
	Lung cancer specific cough			
	Pain in chest		O	
	New possibility			O
	Personal strength		O	O
	Appreciation of life		O	
Activities	Physical activity (MET)		O	O



Table III-18. Selected important variables based on AdaBoost models

Factors	Variables	Model AdaBoost	Model AdaBoost	Model AdaBoost
		-1	-2	-3
Health Condition	Cancer stage	O		
	Regional lymph node metastasis	O		
Environmental Factors	Time since diagnosis	O		
	Low household income	O		
Personal Factors	Age	O		O
	Sex	O		
Body function and structures	BMI(kg/m2) before operation		O	O
	Anxiety		O	O
	Depression		O	O
	Dyspnea		O	O
	Appetite loss		O	O
	Diarrhea			
	Lung cancer specific cough		O	

## ***D. Model validation and performance***

### **1. Cox proportional hazard ratio model internal validation**

#### **1.1. Discrimination**

The discriminatory ability of the Cox model was measured using the C-statistic in both development and validation sets (Table III-13). The C-statistics for model Cox-1 were 0.687 (in the development set) and 0.699 (in the validation set), while the statistics for model Cox-2 were 0.769 (in the development set) and 0.767 (in the validation set). For model Cox-3, the C-statistics were 0.797 (in the development set) and 0.809 (in the validation set). The values for model Cox-3 showed the highest C-statistics, whereas those for model A showed the lowest values. The models' AUC significantly increased from model Cox-1 through model 3, in both development and validation sets. The final AUC value of each development and validation set in Cox-1, Cox-2, and Cox-3 model are graphically shown in Figure III-6 and Figure III-7.

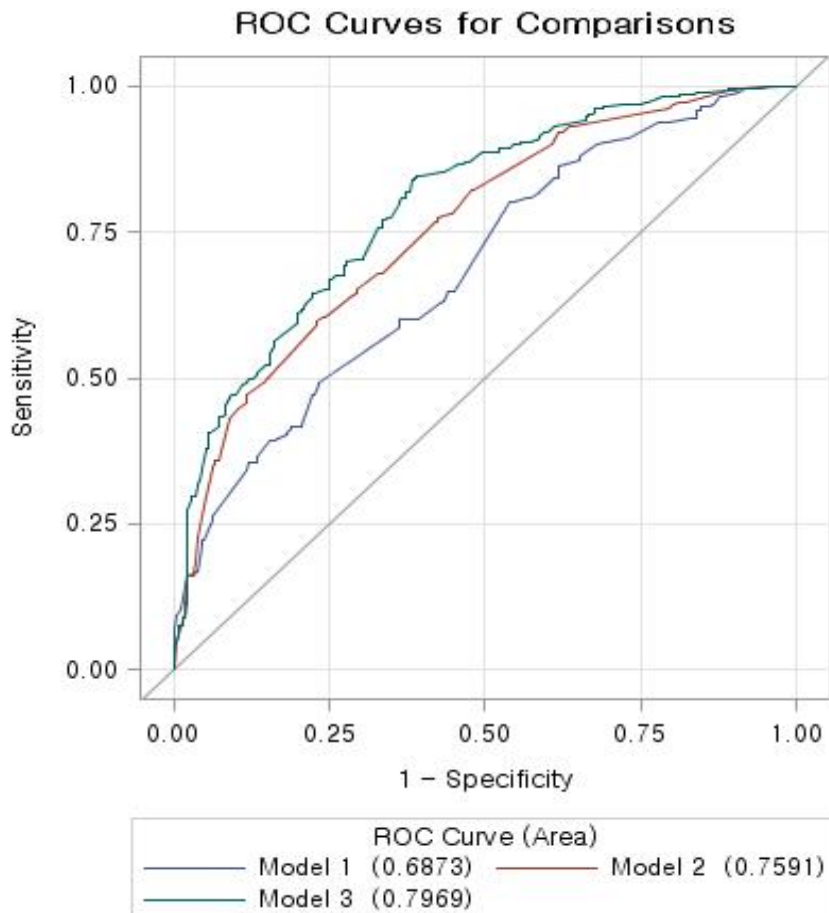


Figure III-6. ROC plot of Cox regression model in development set  
 Model 1 refers to Model Cox-1, Model 2 refers to Model Cox-2,  
 and Model 3 refers to Model Cox-3

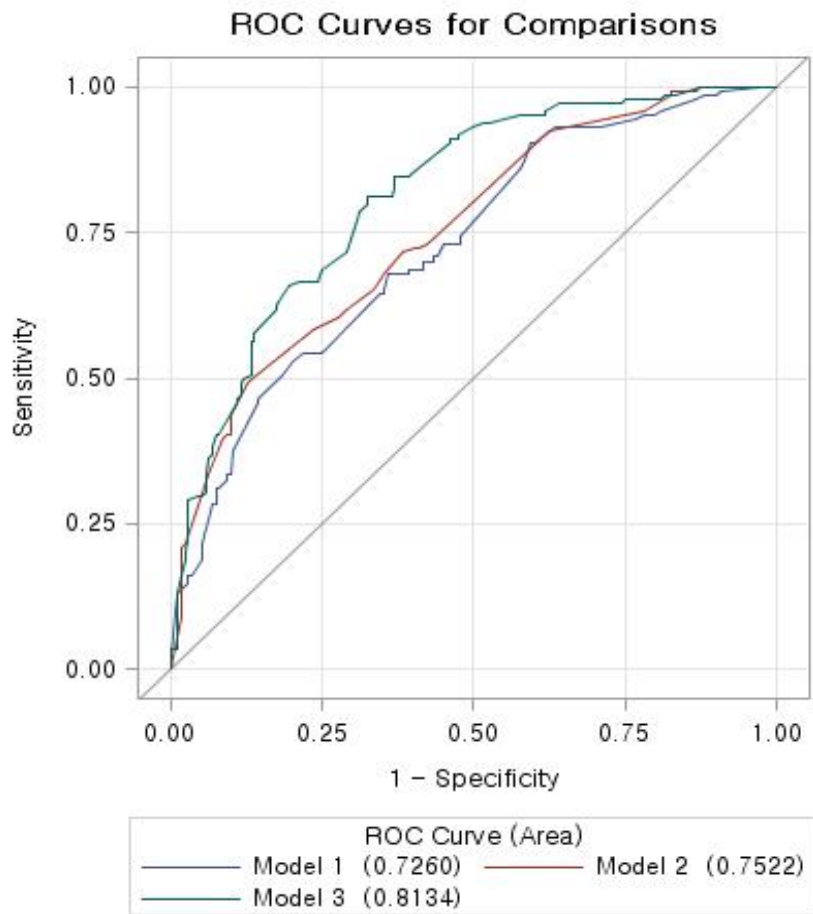


Figure III-7. ROC plot of Cox regression model in validation set  
 Model 1 refers to Model Cox-1, Model 2 refers to Model Cox-2,  
 and Model 3 refers to Model Cox-3

## **1.2. Calibration**

Hosmer-Lemeshow goodness of fit test for models Cox-1, Cox-2 and Cox-3 are shown in Figure III-8~10. The prediction values are plotted in the x-axis and the H statistic, which is based on the fixed cut-points on the predictions, are plotted as decile groups in the calibration graphs. Each figure shows the calibration plots for the overall lung cancer survival prediction model as well the expected/observed (E/O) ratios of validation sets for models Cox-1, Cox-2 and Cox-3. The calibration plot aligns well with the diagonal line. In model Cox-1, calibration in Hosmer-Lemeshow p-value was significantly close to the actual observations ( $p=0.0002$ ). In models Cox-1 and Cox-2 the p value was 0.0019 indicating a good calibration. In model Cox-3, the p-value was 0.0078, also showing a good calibration. In general, the event rates predicted by the models were significantly similar to the actual event rates in all the three models.

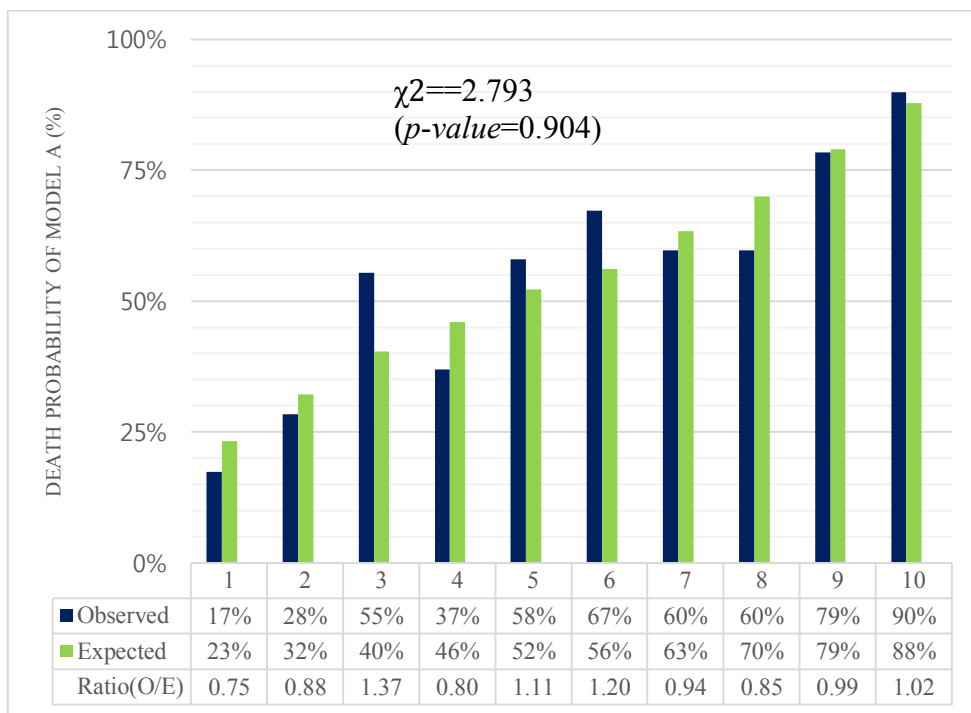


Figure III-8. Calibration plot of lung cancer prediction model Cox-1

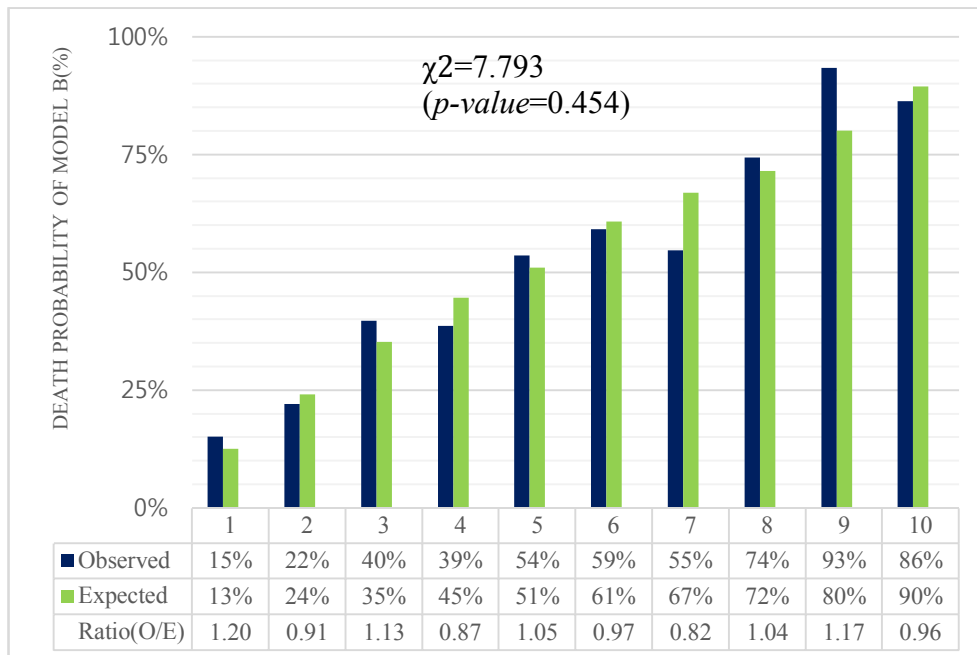


Figure III-9. Calibration plot of lung cancer survival prediction model Cox-2

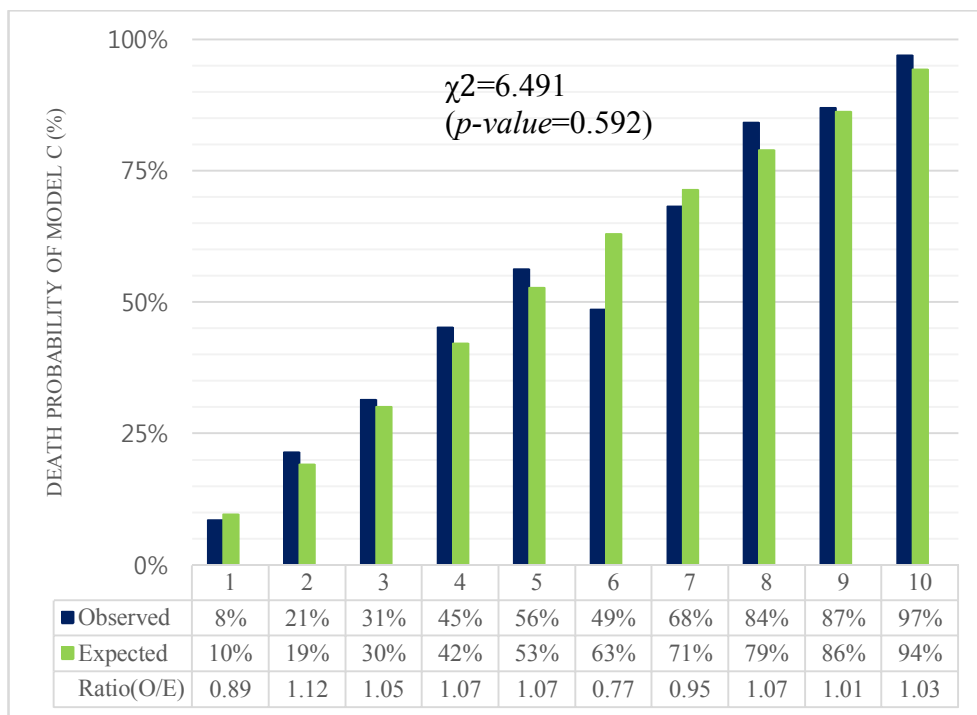


Figure III-10. Calibration of lung cancer prediction model Cox-3

Table III-19. C-statistic and Hosmer-Lemeshow type chi-square test for lung cancer survival prediction models for development and validation sets

	Increase in AUC (p)				
	Model Cox-1	Model Cox-2	Model Cox-3	Model Cox-2 – Model Cox-1	Model Cox-3 - Model Cox-1
Development set					
C (95% CI)	0.687 (0.649-0.725)	0.759 (0.725-0.793)	0.797 (0.765-0.829)	0.072 (0.003)	0.110 (<0.001)
Chi-square value (p-value)	11.883 (0.105)	19.480 (0.0125)	9.571 (0.297)		
Validation set					
C (95% CI)	0.699 (0.668-0.730)	0.767 (0.739-0.795)	0.809 (0.783-0.835)	0.068 (0.001)	0.1102 (<0.001)
Chi-square value (p-value)	2.793 (0.904)	7.793 (0.454)	6.491 (0.592)		



## **2. Comparison model performance of Cox model and other MLTs**

Positive predictive value (PPV), negative predictive value (NPV), sensitivity, specificity, accuracy, and AUC with 95% CI were calculated. On the basis of accuracy and AUC of the prediction model with validation (or test) set, each of the Cox and MLT based prediction model performance were measured. (Table III-20~24)

The overall model performances of ‘Feature set 3’ proved superior to the other feature sets. Among the MLTs, DT showed the lowest performance in both accuracy and AUC, RF showed the highest AUC in models RF-1, RF-2, and RF-C (0.821, 0.789, and 0.918, respectively), while bagging showed the highest accuracy (%) in models Bagging-1 and 2 (73.5% and 74.1%, respectively) and AdaBoost showed the highest accuracy (%) in model AdaBoost-3 (84.9%). In model AdaBoost-1, AdaBoost showed the lowest PPV, while it was the highest in model AdaBoost-3, (84.43). In general, ensemble MLTs, including RF, bagging, and AdaBoost showed better performances.

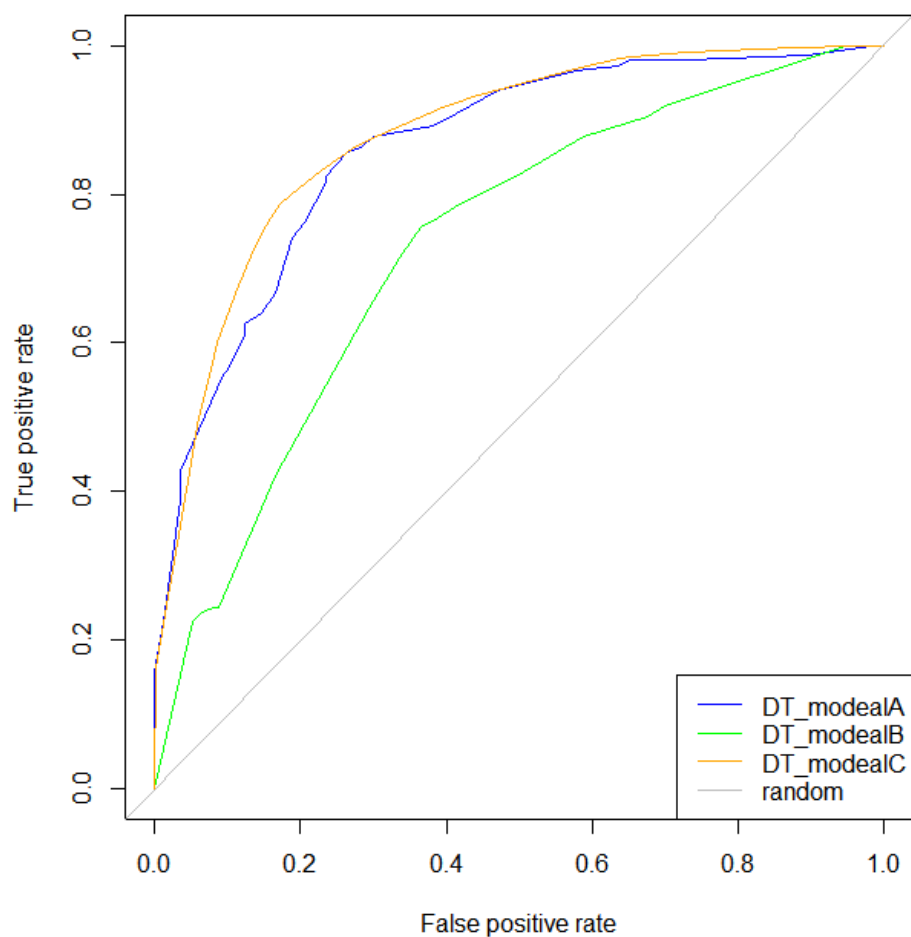


Figure III-11. AUC curve comparison of lung cancer prediction models based on decision tree  
(DT-1: DT\_modelA, DT-2: DT\_modelB, DT-3: DT\_modelC)

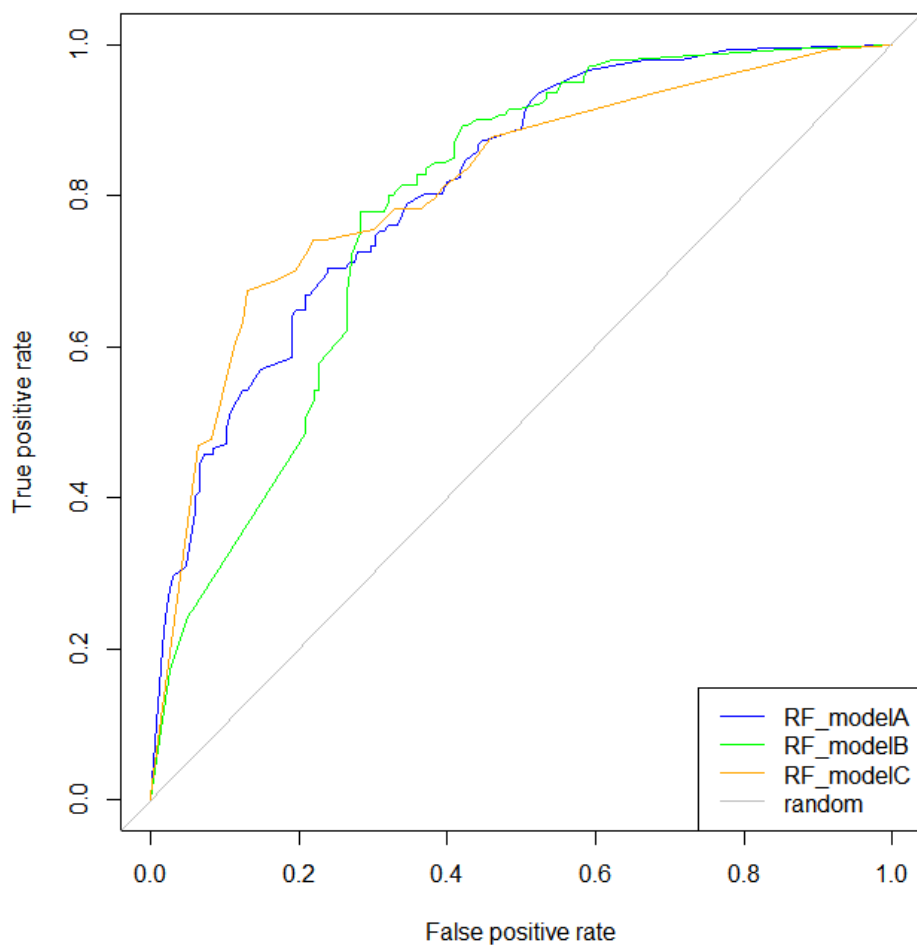


Figure III-12. AUC curve comparison of lung cancer prediction models based on random forest model (RF-1: RF\_modelA, RF-2: RF\_modelB, RF-3: RF\_modelC)

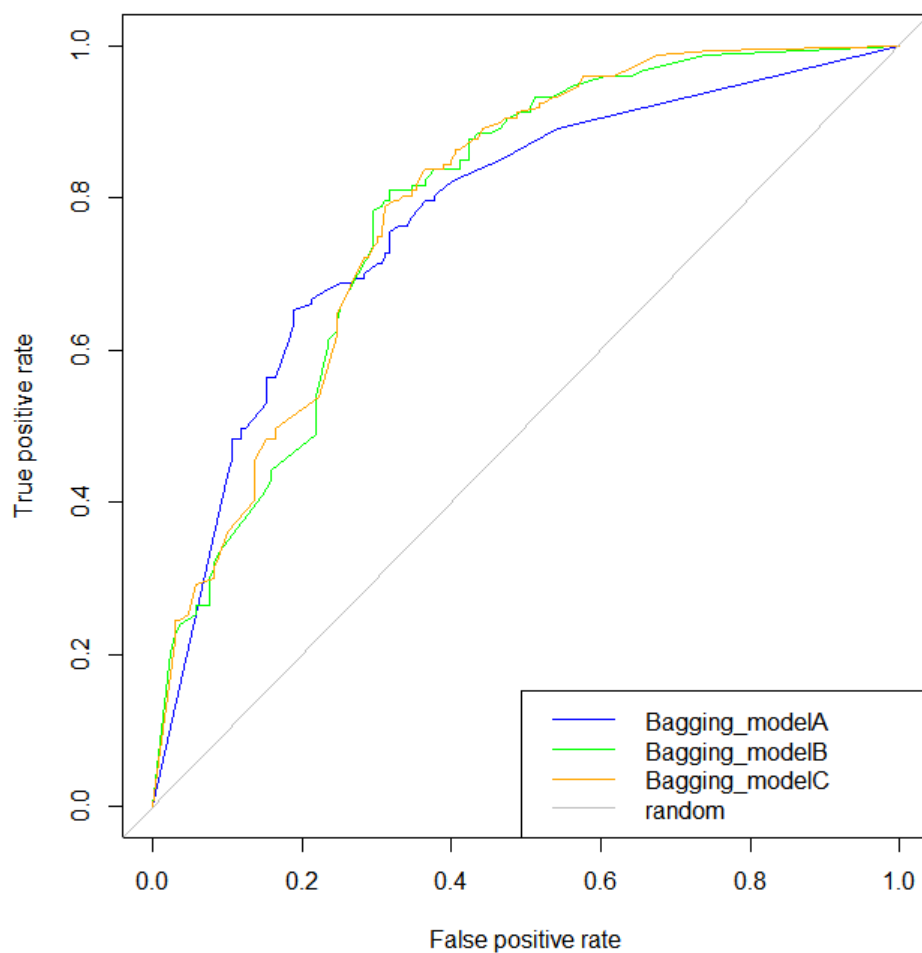


Figure III-13. AUC curve comparison of lung cancer prediction models based on Bagging techniques  
 (Bagging-1: Bagging\_modelA, Bagging-2: Bagging\_modelB, Bagging-3: Bagging\_modelC)

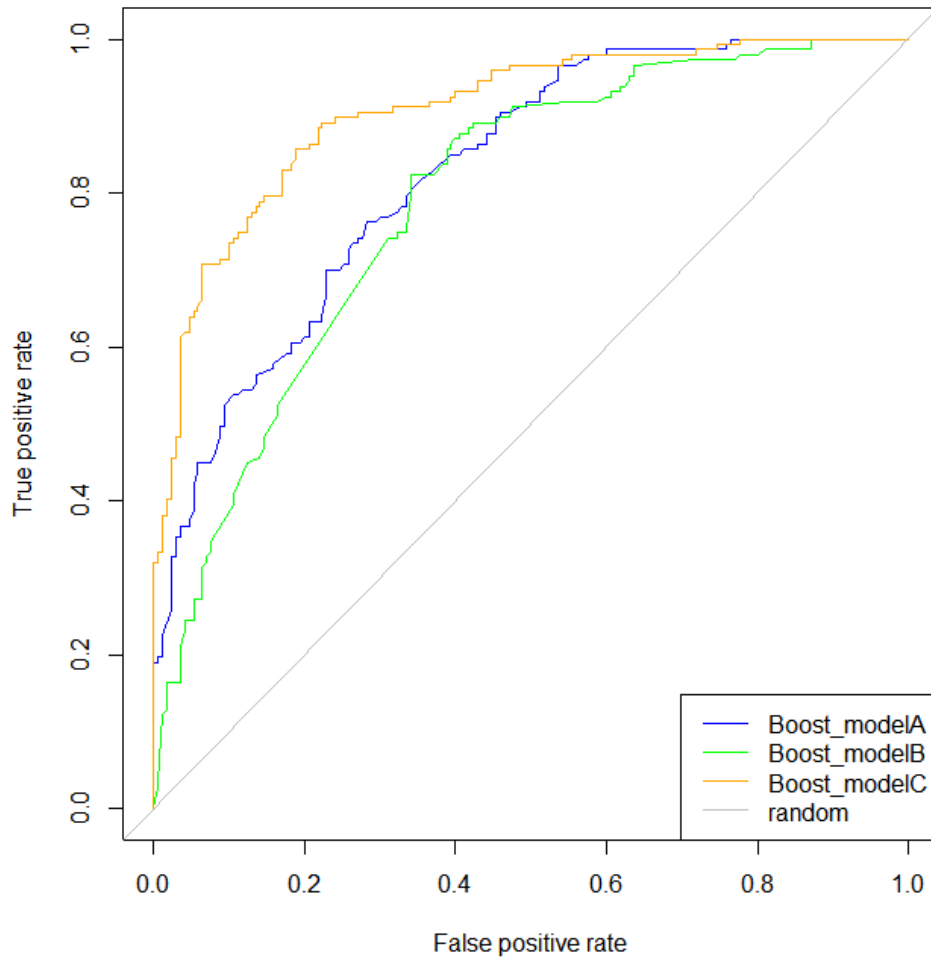


Figure III-14. AUC curve comparison of lung cancer prediction models based on Bagging techniques  
 (AdaBoost-1: Boost\_modelA, AdaBoost-2: Boost\_modelB, AdaBoost-3: Boost\_modelC)

Table III-20. Performance comparison of three data mining algorithms based on the Cox models

	PPV, Precision	NPV	Sensitivity (TPR)	Specificity	Accuracy (%)	AUC (95% CI)
Model Cox-1	72.41	71.96	83.05	57.89	72.30%	0.699(0.668-0.730)
Model Cox-2	70.30	67.26	79.33	55.88	69.20%	0.767(0.739-0.795)
<b>Model Cox-3</b>	<b>76.29</b>	<b>75.00</b>	<b>83.62</b>	<b>65.41</b>	<b>75.80%</b>	<b>0.809(0.783-0.835)</b>

**Abbreviation:** PPV, Positive Predictive Value; NPV, Negative Predictive Value

Table III-21. Performance comparison of three data mining algorithms based on the DT models

	PPV, Precision	NPV	Sensitivity (TPR)	Specificity	Accuracy (%)	AUC (95% CI)
Model DT-1	68.24	72.79	74.36	66.46	69.10%	0.775(0.724-0.826)
Model DT-2	77.70	65.17	63.53	78.91	67.50%	0.769(0.717-0.822)
<b>Model DT-3</b>	<b>75.50</b>	<b>66.27</b>	<b>67.06</b>	<b>74.83</b>	<b>76.30%</b>	<b>0.800(0.750-0.850)</b>

**Abbreviation:** DT, Decision Tree; PPV, Positive Predictive Value; NPV, Negative Predictive Value

Table III-22. Performance comparison of three data mining algorithms based on the RF models

	PPV, Precision	NPV	Sensitivity (TPR)	Specificity	Accuracy (%)	AUC (95% CI)
Model RF-1	73.99	70.80	76.19	68.31	73.50%	0.821(0.775-0.867)
Model RF-2	79.41	68.71	67.92	80.00	73.80%	0.789(0.728-0.830)
<b>Model RF-3</b>	<b>86.16</b>	<b>84.96</b>	<b>87.26</b>	<b>83.70</b>	<b>82.30%</b>	<b>0.918(0.888-0.947)</b>

**Abbreviation:** RF, Random Forest; PPV, Positive Predictive Value; NPV, Negative Predictive Value

Table III-23. Performance comparison of three data mining algorithms based on the Bagging models

	PPV, Precision	NPV	Sensitivity (TPR)	Specificity	Accuracy (%)	AUC (95% CI)
Model Bagging-1	73.53	66.67	71.84	68.53	73.5%	0.788(0.740-0.837)
Model Bagging-2	76.44	71.33	76.44	71.33	74.1%	0.779(0.728-0.830)
<b>Model Bagging-3</b>	<b>81.46</b>	<b>82.01</b>	<b>85.29</b>	<b>77.55</b>	<b>77.9%</b>	<b>0.834(0.789-0.878)</b>

**Abbreviation:** PPV, Positive Predictive Value; NPV, Negative Predictive Value

Table III-24. Performance comparison of three data mining algorithms based on the AdaBoost models

	PPV, Precision	NPV	Sensitivity (TPR)	Specificity	Accuracy (%)	AUC (95% CI)
Model AdaBoost-1	71.351	71.212	77.647	63.946	72.2%	0.819(0.774-0.864)
Model AdaBoost-2	75.497	66.265	67.059	74.830	73.5%	0.785(0.735-0.835)
<b>Model AdaBoost-3</b>	<b>84.431</b>	<b>79.137</b>	<b>82.941</b>	<b>80.882</b>	<b>84.9%</b>	<b>0.893(0.838-0.927)</b>

**Abbreviation:** PPV, Positive Predictive Value; NPV, Negative Predictive Value

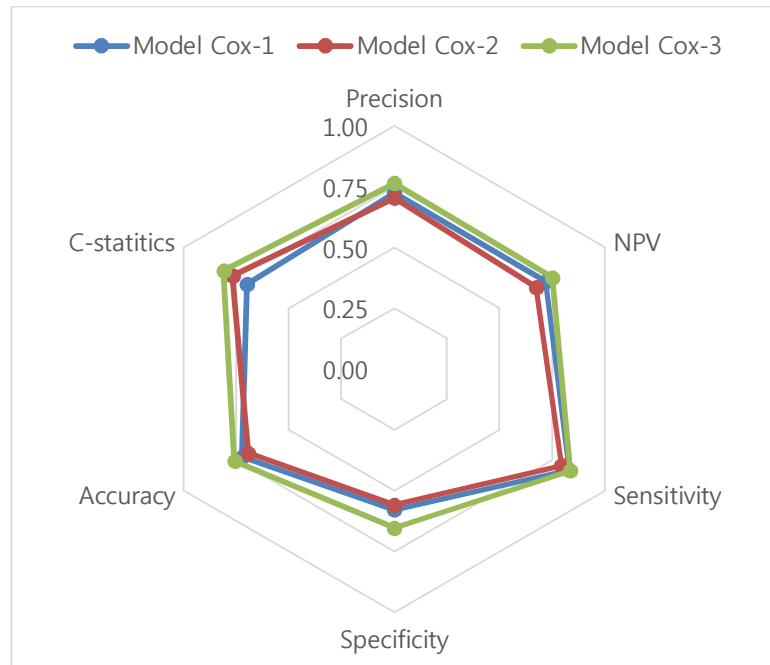


Figure III-15. Rader chart of performance of three models based on Cox model

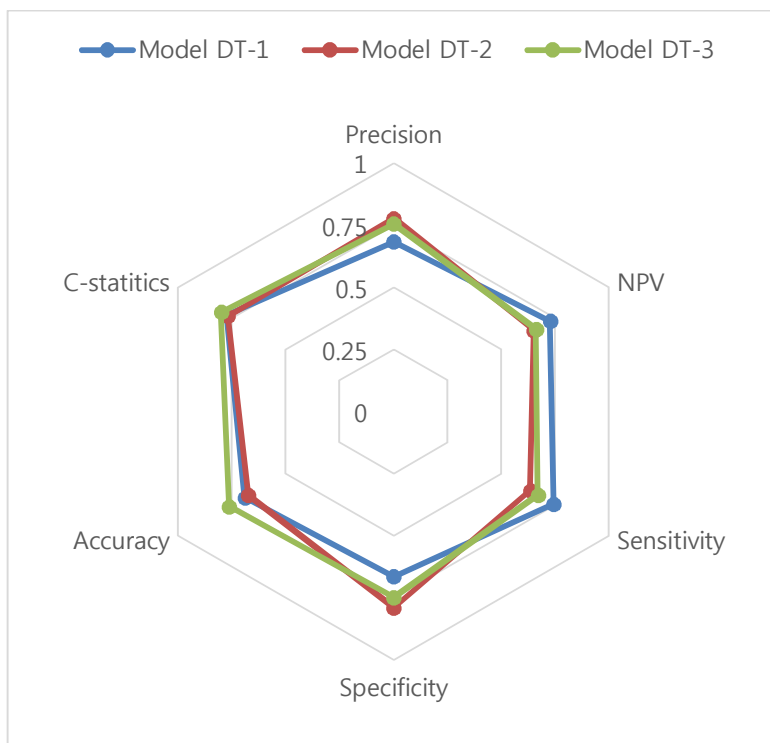


Figure III-16. Rader chart of performance of three models based on DT model



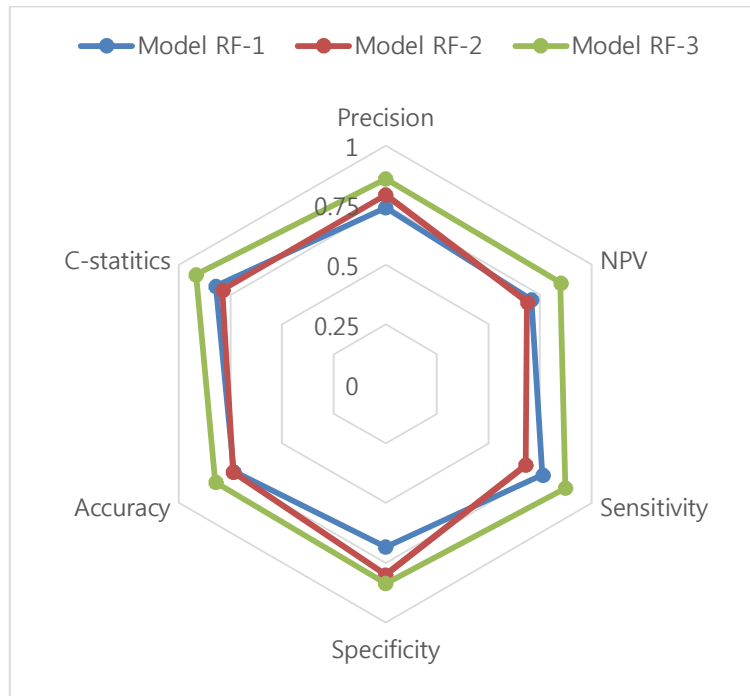


Figure III-17. Rader chart of performance of three models based on RF model

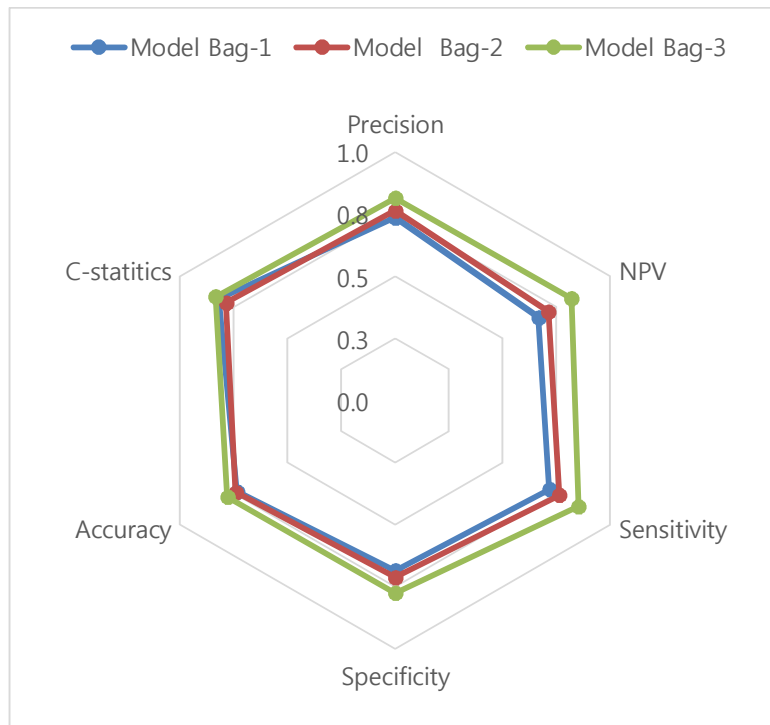


Figure III-18. Rader chart of performance of three models based on bagging model

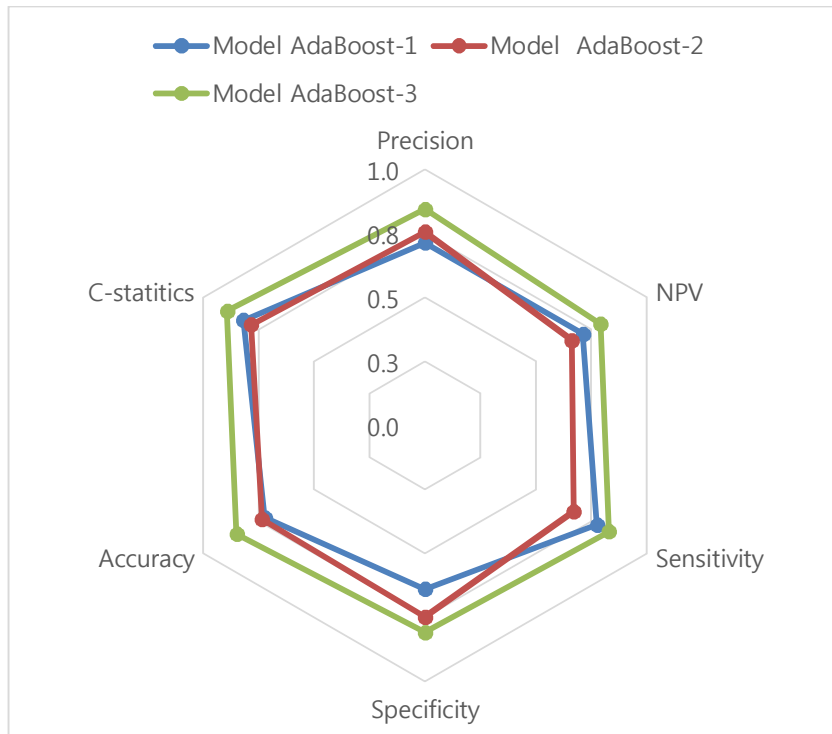


Figure III-19. Rader chart of performance of three models based on AdaBoost model

## IV. Discussion

A goal of this study, which was to construct a novel Korean survival prediction model within lung cancer disease-free survivors using socio-clinical and HRQOL variables and to compare its predictive performance with the prediction model based on the traditional known clinical variables were undertaken and validated using Cox proportional hazard model and machine learning technologies (MLT) mentioned below. In this study, we first assessed the possible prognostic factors of lung cancer mortality through literature review focusing on HRQOL measurements, and selected the factors that contribute to lung cancer survival, subsequent to identifying variables constructing three feature sets. Further, diverse techniques such as Cox proportional hazard model and MLTs were applied to the modeling process and the best models that explain individual prognostic factors of lung cancer survivors' traditional clinical variables, and HRQOL measurements and its interaction in the development of lung cancer survival prediction model were investigated and compared within the same techniques. The model considering clinical and HRQOL measurements together showed the best accuracy for Cox modeling. In addition, validity of the developed mathematical model in lung cancer survival data from feature set 3 (including clinical and HRQOL measurements together) established the best model in each of four MLTs (DT, RF, bagging, AdaBoost), where AdaBoost showed the best model performance compared with other MLTs.

## **A. Literature review for selection of candidate predictors**

Based on our knowledge, there are less mathematical models that have been developed and validated to predict the development of lung cancer mortality based on the HRQOL factors. Thus, selection of candidate prognostic factors and reducing it before studying the prognostic factors–mortality relationship was critical in increasing the robustness and validity of the model built. Because prognostic factor analysis of HRQOL variables is complex, and rarely standardized [86], we tried to select the final modeling candidate prognostic factors based on the systematic process of literature review, not just constructing the prediction models through the feature selection of MLTs. In addition, those of variable selection process can be helpful to avoid over-fitting somewhat. We hope that candidate prognostic factors summarized here will guide the further studies related with prognostic factor analyses. The value of our survival prediction model is on the ability to predict with other lung cancer populations, thus reasonable candidate predictors should be chosen to be generalizable and to also avoid over-fitting in the data that was used for model development.

The strategy of selecting the candidate predictors was based on extensive literature search in the field of lung cancer prognostic factor analyzing studies. In specific, we attempted to find the lung cancer survival prognostic factors focusing on HRQOL measurements and lifestyle factors. Moreover, existing literature reviews of lung cancer survival prognostic factors were sought based on published cross-sectional and cohort studies. The evidence level of some predictors were diverse between those searched by literature. Thus, the study of univariate analysis from statistical analyses complemented our data.

In addition to the well-known traditional assessment of clinical or socio-demographic characteristics, HRQOL outcomes may have other applications that include supporting clinical decision-making by providing cancer survivors' prognostic information [87]. Considering one of the HRQOL studies which reports psychosocial well-being as a prognostic factor of survival in non-small-cell lung cancer patients [88], HRQOL prognostic factor analysis studies started to examine diverse HRQOL questionnaires with different cancer populations [39, 89, 90].

From literature review, we found stronger rationales for HRQOL selection as lung cancer prognostic factors in the survival prediction model development. From the study review, excluding few exceptions, the findings showed that QOL data or HRQOL assessments were significant independent predictors of survival duration. Global QOL, functioning domains, and symptom scores, such as appetite loss, fatigue, and pain or dyspnea, were the most predictive indicators of lung cancer survivors, individually or in combination, in predicting survival times in cancer survivors after adjusting for one or more socio-demographic and known clinical or medical prognostic factors [90].

This systematic review provides proficient evidence for a positive association between some HRQOL assessments and the survival duration of cancer survivors. Baseline QOL data or life style factors appeared to provide the most reliable information for helping clinicians to establish prognostic criteria for long-term care of cancer survivors.[91] It is recommended that future studies should use valid instruments and apply sound methodological approaches.

However, in addition to considering HRQOL in the survival prediction model, subjects which were significant in adequate multivariate statistical analyses

adjusted for socio-demographic characteristics and clinical prognostic factors from literature reviews and previous analysis should be considered. The current systematic review results demonstrate that for lung cancer survival, HRQOL functions and symptoms provided prognostic information when those factors were additive over clinical or socio-demographic characteristics [92]. This strategy is expected to yield more accurate and specific QOL-related prognostic variables for specific cancers.

Finally, five socio-demographic and clinical values were selected from literature review, a total of 13 variables of HRQOL, and 2 lifestyle factors were chosen before the development of the model. Non only considering previous studies, to achieve a stronger evidence which would better explain or be significant in our data, we selected the variables which were shown as significant in our univariate analysis or previous prognostic factors used to analyze the same data [91].

## **B. Model development using Cox and other MLTs**

In survival analysis, several different regression modeling techniques can be applied to predict the prognostic factors of an event occurring. However, very often, the default choice may rely on Cox regression proportional modeling due to its convenience. In this situation, extensions of the machine learning algorithms to survival analysis provide an alternative approach to build a more accurate survival prediction model [76]. In general, all types of prediction models can be investigated, where diverse packages readily support traditional Cox regression or hazard regression, as well as state of the art machine learning methods, including

ensemble modeling methods which provide promising alternatives to traditional strategies in both low and high-dimensional settings [93].

Therefore, in our study, each of the Cox models developed and other four modeling techniques were also used to seek the best fitting survival prediction model for lung cancer survivors. For the MLTs, decision survival tree, and three ensemble learning, including RF model, bagging, and AdaBoost were used. Each of the three models 1) model from clinical and socio-demographic variables, 2) HRQOL prognostic factors model, and 3) clinical and socio-demographic variables and HRQOL factors combined model, were developed using the development set. Cox regression and other MLTs which were applied in our study are widely used mathematical techniques used to develop supervised classification models. Studies have compared the discriminatory ability of the 3 models based on five techniques and to identify the key covariates to predict the survival outcome.

The MLTs which we chose have the key concepts of supervised learning, with several advantages, such as lowering training loss resulting in a more predictive model, and lowering regularization resulting in a simpler model. A DT is a decision support tool that uses a tree-like graph, including chance event outcomes, resource costs, and utility [72, 78]. Every ensemble method is a learning technique that constructs a set of classifiers and then classifies new data points by taking a weighted vote of their predictions, starting from the case of DTs [94, 95]. This procedure leads to better model performance as it decreases the variance of the model, without increasing the bias.

RFs use a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features. This

process is sometimes called “feature bagging.” The reason for doing this is the fact that the correlation of trees in an ordinary bootstrap sample: if one or a few features are very strong predictors for the response variable (target output) then these features will be selected in many of the trees, causing a correlation.

While boosting is not algorithmically constrained, most boosting algorithms consist of iteratively learning weak classifiers with respect to a distribution and add them to a final strong classifier. After a weak learner is added, the data are reweighted: examples that are misclassified gain weight and examples that are classified correctly lose. Thus, future weak learners focus more on the examples that previous weak learners misclassified.

In our study, to develop a simple and robust prediction model based on diverse mathematical technologies that can be readily used in a real life setting, all the continuous variables were categorized according to the clinically meaningful cut-offs. Perhaps, as standardized scores of the HRQOL that range from 0 to 100 would not require complex mathematical formulae to explain the independent and dependent binary variables, we used the cut-off points based on the stronger rationale, which refer to the problematic groups in the real medical setting.

For non-linear modeling, classification MLTs, such as DT model or several ensemble algorithms, were proposed as one of supplement to the Cox regression [76, 79, 95-99]. However, because of the difficulty of handling the time of event, we may propose the MLTs application as only as an auxiliary means. Finally, the present study found that even if the model performance is different to the model construction, DT is not better than other MLTs, while other ensemble algorithms, such as RF, bagging, and AdaBoost maybe give more accurate



information or suggest better model performance according to additive HRQOL variable selection to clinical variables.

### **C. Model validation of Cox regression model and application of the predictive models to other MLT based models**

In this study, based on the disease-free lung cancer survivors' data, we intended to test two hypotheses that if the survival prediction model is analyzed by diverse techniques (i.e., Cox regression model or MLTs) and if the model includes not only traditional clinical or socio-demographic variables, but also patient report outcome assessment such as survivors' HRQOL or lifestyle factors, the routinely collected health data through survivor's report could be used to assist conventional tools in predicting clinical outcomes, and monitor survivors' medical status in comparison to traditional prediction model development, such as clinical data based prediction modeling.

Therefore, we first compared and validated the prognostic accuracy of the Cox regression model, where a feature set 1) model only selected clinical or socio-demographic variables, 2) model contained HRQOL factors, and 3) model combined all variables. Further, model performance was also applied to models based on the MLTs. Finally, the application of RF or AdaBoost to data from the model integrating HRQOLs or lifestyle factors with traditional well-known clinical variables predicted lung cancer survival with the best accuracy or AUC in MLT based models.

The Cox model showed moderately good discrimination and well

calibrated performance; however, several ensemble learning (i.e., RF, bagging, and AdaBoost) showed significantly greater model AUC with 95% CI and proved superior in terms of accuracy (%). Similar to a previous study [96], our lung cancer survival prediction model was developed with not only the clinical or socio-demographic factors, but also model integration on information from multiple factors, such as HRQOL or lifestyle factors additive to clinical factors, which ensured a better model performance and even greater predictive ability in all modeling techniques. Among several lung cancer survival prediction models developed and validated, most studies focus on a specific group of variables independently (i.e., socio-demographic variables, biomarkers, genetic information, imaging, or clinical characteristics). Therefore, integrating information from multiple sources, such as patient reported outcome with clinical and treatment variables in model developing is the way to move forward toward more accurate prediction models [96].

In the statistical approach, after the model is inferred, the process of verifying the statistical significance of the model is performed. Machine learning is done in the direction of optimizing accuracy indicators such as prediction accuracy, MSE, etc. when making a steel model. As a result, many machine learning models are black-boxed and judging how good or bad the model is according to how much predictive performance is. When the size of data is very large, various information is mixed in it, it is difficult to find a pattern in it, it is not easy to review the assumption of data when performing statistical modeling. Since it is important to improve prediction accuracy, it is meaningful to make a survival prediction model using machine learning techniques.

In our study, the model performance for ensemble learning such as random forest or AdaBoost also applied into the prediction model. Unfortunately, limited survival prediction models from previous studies used diverse MLTs, including ensemble techniques and compared performances. There were several trials to compare SVM- or ANN-based breast cancer recurrence prediction models with Cox [18, 100], or those that attempted to predict cancer survivors' HRQOL itself or DT used in prostate cancer survival [97]. However, it was challenging to find a comparative study of models developed based on the ensemble and to find a model considering HRQOL factors altogether. The ensemble techniques are well established in the field of machine learning, but are almost completely unknown as a lung cancer predictive and prognostic method. In addition, the MLTs we used into modeling process were not appropriate to predict the time of occurrence of an event, there can be limitation. Even though, no lung cancer prognostic model based on clinico-pathologic data and HRQOL altogether have been developed using MLTs [18]. It provides us with a mathematical understanding of the inputs, for which the learning method was employed.

#### **D. Clinical and practical implications**

There are several possible explanations for the findings of this study. First, lung cancer survivors' HRQOL information played a key role in survival, considering traditional assessments of clinical outcomes. From systematic review, we found that there were impressive number of studies with a positive association between HRQOL and survival. Based on the supportive background, we developed three kinds of Cox proportional hazards models, which included both clinical variables

and HRQOL factors that quantified better predictive accuracy in our data.

Even if biomedical or clinical parameters are generally known as the first factors having prognostic values, HRQOL parameters have been regarded as added values in predicting survival. Indeed, our data included HRQOL data in the prediction model and showed a better prediction of survival duration when added to traditional clinical and socio-demographic variables. Several previous studies computed discriminative C-indexes of HRQOL's other cancer survival prediction models. They found that the increasing C-indexes were observed when the HRQOL variables were added to the initial clinical variables. However, for lung cancer, this is the first study that compares the accuracy of HRQOL included prediction model with traditional models.

Second, our study has suggested the meaning of HRQOL prediction model toward cancer survivors' self-management. The prognosis of disease-free survivors with non-small cell lung cancer is significant for both clinical and basic research [101]. The identification of prediction model can help providing information for cancer survivors, as well as aid physicians in choosing the best method for surveillance and intervention. However, less previous studies have addressed the prediction model based with HRQOL for disease-free lung cancer survivors who are regarded to have a relatively good prognosis after the completion of active cancer treatment. It is also possible that individuals with poor HRQOL, or those who are not motivated, may be less likely to adhere to their medical treatment plans [102] and good health behaviors (such as moderate-to-vigorous PA) that are independent predictors of mortality in disease-free lung cancer survivors [35, 103, 104].

Even if we cannot change the clinical factors, HRQOL or lifestyle factors can be modified. In our study, dyspnea, personal strength, BMI, Physical activity anxiety and depression were selected as important variables from diverse modeling. Additionally, it is possible that role functioning [86, 105-107], dyspnea [38, 91, 108-111], fatigue [86, 106, 109, 112], cough [109], anxiety [91, 110] and depression [91, 110, 111] are strong prognostic variables for survival in advanced lung cancer during clinical trials or after the treatment.[38, 113] Post-traumatic growth factors were also known as good prognostic value in disease-free lung cancer survivors.[91] These findings may indicate a disease progression or recurrence that physical examination by a clinician, tumor marker evaluation, and imaging studies (such as computed tomography, magnetic resonance imaging, and positron emission tomography) could not detect.[35] Therefore, better lung cancer prognostic indexes need to be developed based on both clinical and HRQOL factors, and we need to develop the individual assessment algorithms of prognosis of survivors, guiding the clinical decision-making system to provide more information about their care.

Finally, this study has shown that the ability to apply new prediction algorithms based on diverse machine learning. Even though MLTs have been used to analyze gene expression data studies [114-116] or medical image prediction analyses [117], the studies which explored MLTs in clinical settings were not sufficient. Gradually, there have been several trials to ensure more sophisticated and better validated techniques, and the need to improve model accuracy to a reasonable level. In our study, the approach used offers superior performance in comparison to previous machine-learning approaches in predicting cancer survivals.

In addition, they could be used to better stratify lung cancer survivors in future cancer clinical trials; therefore, improving the interpretation of study outcomes or helping identify critical areas that could help in the selection of key end points for future clinical trials [46]. Despite superior performance of machine learning algorithms, the use of such algorithms in daily clinical practice has been rather limited as they cannot be easily calculated with a traditional calculator; thus, MLT algorithms could be the black box in a sense [18, 118, 119]. Therefore, for the convenience of usage in clinical settings, developing a comprehensive ICT self-management program by including the prediction model can provide more information and help survivors' decisional support [16].

## **E. Strengths and limitations of this study**

Although several studies predict lung cancer survival based on the MLTs, based on our knowledge, this is the first study that used HRQOL factors with clinical and socio-demographic variables to develop a lung cancer survival prediction model. On the other hand, previous research models were derived from traditional clinical variables, and the models described here worked well in models that consider HRQOL variables together. A well-made and low-cost lung cancer prediction model can be implemented into the ICT-based self-management care system and could help patients ensure improving their HRQOL as well as their satisfaction for new paradigm of health care model[18] In the further study, based on the developed prediction model, we may apply the models into the web or app programs and investigate the effectiveness of the prediction model through the RCT studies.

Our study may be noted for several limitations. First, there could be a selection bias. As only disease-free lung cancer patients from two of the big hospitals who survived at least 1 year after surgery participated in this study, our sample could not represent general lung cancer patients, meaning that generalization of our findings to similar groups of cancer patients may be restricted. Second, this study only addressed overall mortality and did not include lung cancer-specific mortality and non-cancer mortality. Further studies that include cancer-specific mortality and non-cancer mortality would be helpful for interpreting the prognostic value of HRQOL in lung cancer. Third, the machine learning techniques adapted to effectively handle survival data should be investigated.[48] However, the MLTs which we applied to this study cannot accurately predicting the time of occurrence of an event. To handle survival problems with machine learning algorithms, more effective algorithms incorporating survival problems with both statistical methods and machine learning techniques such as survival trees[120], random survival forests[121], bagging survival trees[79] and boosting[122]. Therefore, in the further study we need to develop the MLTs which we can handle survival problems. Finally, the participants were surveyed at different time intervals from the time of their diagnosis. Thus, we adjusted for this as a co-variable investigating time since diagnosis more than 3 years. Even if we suggest the assessment of HRQOL and those of lung cancer prediction model based on the prognostic factors to be incorporated into routine oncology clinical practice, further studies, such as randomized controlled trials should be conducted and the efficacy of a prediction model based program should be validated.

## **CONCLUSIONS**

In conclusion, consideration of PRO information (including HRQOLs and life style factors) added to clinical and socio-demographic variables in lung cancer survival prediction model proved to be more accurate than traditional clinico-pathological variables based Cox prediction model. In addition, we found that the same proposed feature set can be applied into ensemble MLT algorithms (particularly random forest or Adaboost algorithms) to predict disease-free lung cancer survival. Most importantly, considering HRQOL and lifestyle factors together in a lung cancer survival prediction process will suggest patients more accurate information and lower their decisional conflicts. Improved accuracy for lung cancer survival prediction model has the potential to help clinicians and survivors make more meaningful decisions about future plans and their support to cancer care.



## REFERENCES

1. Jung, K.W., et al., *Prediction of cancer incidence and mortality in Korea, 2012*. Cancer Res Treat, 2012. **44**(1): p. 25-31.
2. Hong Gwan Seo, J.H.P., So Young Kim, Hyung Kook Yang, Eun Joo Nam, *Cancer Facts & Figures 2013 in the Republic of Korea*. 2013, National Cancer Center, Ministry of Health and Welfare. : Ilsan, Republic of Korea.
3. Jung, K.W., et al., *Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2009*. Cancer Res Treat, 2012. **44**(1): p. 11-24.
4. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2015*. CA: a cancer journal for clinicians, 2015. **65**(1): p. 5-29.
5. Torre, L.A., et al., *Global cancer statistics, 2012*. CA Cancer J Clin, 2015. **65**(2): p. 87-108.
6. Rathor, M.Y., et al., *Attitudes toward Euthanasia and Related Issues among Physicians and Patients in a Multi-cultural Society of Malaysia*. J Family Med Prim Care, 2014. **3**(3): p. 230-7.
7. Jung, K.W., et al., *Prediction of Cancer Incidence and Mortality in Korea, 2017*. Cancer Res Treat, 2017. **49**(2): p. 306-312.
8. Kenny LW, H., R.H., & Bryant, C.X. ACSM's Guidelines for Testing and Prescription. Baltimore: Williams and Wilkins 1995; 5th ed.
9. Jung, K.W., et al., *Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2012*. Cancer Res Treat, 2015. **47**(2): p. 127-41.
10. Shin, A., et al., *Lung Cancer Epidemiology in Korea*. Cancer Res Treat, 2017. **49**(3): p. 616-626.
11. Jung, K.-W., et al., *Cancer Statistics in Korea: Incidence, Mortality, Survival, and Prevalence in 2014*. Cancer Research and Treatment : Official Journal of Korean Cancer Association, 2017. **49**(2): p. 292-305.
12. Coups, E.J., et al., *Correlates of physical activity among lung cancer survivors*. Psychooncology, 2009. **18**(4): p. 395-404.
13. Yun, Y.H., et al., *Needs regarding care and factors associated with unmet needs in disease-free survivors of surgically treated lung cancer*. Ann Oncol, 2013. **24**(6): p. 1552-9.

14. McCorkle, R., et al., *Self-management: Enabling and empowering patients living with cancer as a chronic illness*. CA Cancer J Clin, 2011. **61**(1): p. 50-62.
15. Harley, C., et al., *Defining chronic cancer: patient experiences and self-management needs*. BMJ Support Palliat Care, 2012. **2**(3): p. 248-55.
16. Sim, J.A., et al., *Perceived needs for the information communication technology (ICT)-based personalized health management program, and its association with information provision, health-related quality of life (HRQOL), and decisional conflict in cancer patients*. Psychooncology, 2017.
17. Yun, Y.H., et al., *Health-related quality of life in disease-free survivors of surgically treated lung cancer compared with the general population*. Ann Surg, 2012. **255**(5): p. 1000-7.
18. Kim, W., et al., *Development of novel breast cancer recurrence prediction model using support vector machine*. J Breast Cancer, 2012. **15**(2): p. 230-8.
19. Svensson, C.-M., R. Hübner, and M.T. Figge, *Automated classification of circulating tumor cells and the impact of interobserver variability on classifier training and performance*. Journal of immunology research, 2015. **2015**.
20. Hoang, T., et al., *Clinical model to predict survival in chemo-naïve patients with advanced non-small-cell lung cancer treated with third-generation chemotherapy regimens based on Eastern Cooperative Oncology Group data*. Journal of Clinical Oncology, 2005. **23**(1): p. 175-183.
21. Brown, N.M., et al., *Supportive care needs and preferences of lung cancer patients: a semi-structured qualitative interview study*. Support Care Cancer, 2015. **23**(6): p. 1533-9.
22. Philip, E.J., et al., *Physical activity preferences of early-stage lung cancer survivors*. Support Care Cancer, 2014. **22**(2): p. 495-502.
23. Hung, R., et al., *Fatigue and functional impairment in early-stage non-small cell lung cancer survivors*. J Pain Symptom Manage, 2011. **41**(2): p. 426-35.

24. Feinstein, M.B., et al., *Current dyspnea among long-term survivors of early-stage non-small cell lung cancer*. J Thorac Oncol, 2010. **5**(8): p. 1221-6.
25. Payne, C., et al., *Exercise and nutrition interventions in advanced lung cancer: a systematic review*. Curr Oncol, 2013. **20**(4): p. e321-37.
26. Levin, R.M., CSO, LD, *Nutrition in the Patient with Lung Cancer*. Lung Cancer Choices. Vol. 99. 2012.
27. Velikova, G., et al., *Measuring quality of life in routine oncology practice improves communication and patient well-being: a randomized controlled trial*. J Clin Oncol, 2004. **22**(4): p. 714-24.
28. Kao, S.C., et al., *Health-related quality of life and inflammatory markers in malignant pleural mesothelioma*. Support Care Cancer, 2013. **21**(3): p. 697-705.
29. Maione, P., et al., *Pretreatment quality of life and functional status assessment significantly predict survival of elderly patients with advanced non-small-cell lung cancer receiving chemotherapy: a prognostic analysis of the multicenter Italian lung cancer in the elderly study*. J Clin Oncol, 2005. **23**(28): p. 6865-72.
30. Efficace, F., et al., *Is a patient's self-reported health-related quality of life a prognostic factor for survival in non-small-cell lung cancer patients? A multivariate analysis of prognostic factors of EORTC study 08975*. Ann Oncol, 2006. **17**(11): p. 1698-704.
31. Qi, Y., et al., *Pretreatment quality of life is an independent prognostic factor for overall survival in patients with advanced stage non-small cell lung cancer*. J Thorac Oncol, 2009. **4**(9): p. 1075-82.
32. Rodriguez, A.M., N.E. Mayo, and B. Gagnon, *Independent contributors to overall quality of life in people with advanced cancer*. Br J Cancer, 2013. **108**(9): p. 1790-800.
33. Cormio, C., et al., *Post-traumatic Growth in the Italian Experience of Long-term Disease-free Cancer Survivors*. Stress Health, 2015. **31**(3): p. 189-96.
34. Wettergren, L., et al., *Determinants of health-related quality of life in long-*

- term survivors of Hodgkin's lymphoma*. Qual Life Res, 2004. **13**(8): p. 1369-79.
35. Gotay, C.C., et al., *The prognostic significance of patient-reported outcomes in cancer clinical trials*. J Clin Oncol, 2008. **26**(8): p. 1355-63.
  36. Montazeri, A., *Quality of life data as prognostic indicators of survival in cancer patients: an overview of the literature from 1982 to 2008*. Health Qual Life Outcomes, 2009. **7**: p. 102.
  37. Karvonen-Gutierrez, C.A., et al., *Quality of life scores predict survival among patients with head and neck cancer*. J Clin Oncol, 2008. **26**(16): p. 2754-60.
  38. Movsas, B., et al., *Quality of life supersedes the classic prognosticators for long-term survival in locally advanced non-small-cell lung cancer: an analysis of RTOG 9801*. J Clin Oncol, 2009. **27**(34): p. 5816-22.
  39. Efficace, F., et al., *Baseline health-related quality-of-life data as prognostic factors in a phase III multicentre study of women with metastatic breast cancer*. Eur J Cancer, 2004. **40**(7): p. 1021-30.
  40. Rock, C.L., et al., *Nutrition and physical activity guidelines for cancer survivors*. CA Cancer J Clin, 2012. **62**(4): p. 243-74.
  41. Meyerhardt, J.A., et al., *Impact of physical activity on cancer recurrence and survival in patients with stage III colon cancer: findings from CALGB 89803*. J Clin Oncol, 2006. **24**(22): p. 3535-41.
  42. Ibrahim, E.M. and A. Al-Homaidh, *Physical activity and survival after breast cancer diagnosis: meta-analysis of published studies*. Med Oncol, 2011. **28**(3): p. 753-65.
  43. Schmid, D. and M.F. Leitzmann, *Association between physical activity and mortality among breast cancer and colorectal cancer survivors: a systematic review and meta-analysis*. Annals of Oncology, 2014 Mar 18. **Epub ahead of print**.
  44. Frank B. Hu WCW, T.L., Meir J. Stampfer, Graham A. Colditz, JoAnn E. Manson, *Adiposity as compared with physical activity in predicting mortality among women*. New England Journal of Medicine, 2004. **351**:2694-703.

45. Buffart, L.M., et al., *Evidence-based physical activity guidelines for cancer survivors: current guidelines, knowledge gaps and future research directions*. Cancer Treat Rev, 2014. **40**(2): p. 327-40.
46. Mauer, M., et al., *The prognostic value of health-related quality-of-life data in predicting survival in glioblastoma cancer patients: results from an international randomised phase III EORTC Brain Tumour and Radiation Oncology Groups, and NCIC Clinical Trials Group study*. British Journal of Cancer, 2007. **97**(3): p. 302-307.
47. Tredan, O., et al., *Validation of prognostic scores for survival in cancer patients beyond first-line therapy*. BMC Cancer, 2011. **11**: p. 95.
48. Wang, P., Y. Li, and C.K. Reddy, *Machine learning for survival analysis: A survey*. arXiv preprint arXiv:1708.04649, 2017.
49. Cabitza, F., R. Rasoini, and G.F. Gensini, *Unintended consequences of machine learning in medicine*. Jama, 2017. **318**(6): p. 517-518.
50. Consultation, W.E., *Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies*. Lancet, 2004. **363**(9403): p. 157-63.
51. Aaronson, N.K., et al., *The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology*. J Natl Cancer Inst, 1993. **85**(5): p. 365-76.
52. Fayers PM, A.N., Bjordal K et al, *The EORTC QLQ-C30 Scoring manual, 3rd edn*. European Organization for Research and Treatment of Cancer, Brussels, 2001.
53. Zigmond, A.S. and R.P. Snaith, *The hospital anxiety and depression scale*. Acta Psychiatr Scand, 1983. **67**(6): p. 361-70.
54. Bjelland, I., et al., *The validity of the Hospital Anxiety and Depression Scale. An updated literature review*. J Psychosom Res, 2002. **52**(2): p. 69-77.
55. Tedeschi, R.G. and L.G. Calhoun, *The Posttraumatic Growth Inventory: measuring the positive legacy of trauma*. J Trauma Stress, 1996. **9**(3): p. 455-71.

56. Yang, P., *Epidemiology of lung cancer prognosis: quantity and quality of life*. Cancer Epidemiology, 2009: p. 469-486.
57. Clark, M.M., et al., *Motivational readiness for physical activity and quality of life in long-term lung cancer survivors*. Lung Cancer, 2008. **61**(1): p. 117-22.
58. Cieza, A., J. Bickenbach, and S. Chatterji, *The ICF as a conceptual platform to specify and discuss health and health-related concepts*. Das Gesundheitswesen, 2008. **70**(10): p. e47-e56.
59. Cieza, A. and G. Stucki, *The International Classification of Functioning Disability and Health: its development process and content validity*. Eur J Phys Rehabil Med, 2008. **44**(3): p. 303-313.
60. Ware, J.E., *Conceptualization and measurement of health-related quality of life: comments on an evolving field*. Archives of physical medicine and rehabilitation, 2003. **84**: p. S43-S51.
61. Bours, M.J., et al., *Candidate predictors of health-related quality of life of colorectal cancer survivors: a systematic review*. The oncologist, 2016: p. theoncologist. 2015-0258.
62. Hayden, J.A., et al., *Assessing bias in studies of prognostic factors*. Annals of internal medicine, 2013. **158**(4): p. 280-286.
63. Huguet, A., et al., *Judging the quality of evidence in reviews of prognostic factor research: adapting the GRADE framework*. Systematic reviews, 2013. **2**(1): p. 71.
64. Stucki, G., *ICF linking rules: an update based on lessons learned*. J rehabil med, 2005. **37**(37): p. 212-8.
65. Cieza, A., et al., *Refinements of the ICF Linking Rules to strengthen their potential for establishing comparability of health information*. Disabil Rehabil, 2016: p. 1-10.
66. Karim, M.N., et al., *Missing Value Imputation Improves Mortality Risk Prediction Following Cardiac Surgery: An Investigation of an Australian Patient Cohort*. Heart, Lung and Circulation, 2017. **26**(3): p. 301-308.
67. Viswesvaran, C., *Multiple Regression in Behavioral Research: Explanation and Prediction*. Personnel Psychology, 1998. **51**(1): p. 223.

68. Han, K.H., et al., *Factors associated with depression in disease-free stomach cancer survivors*. J Pain Symptom Manage, 2013. **46**(4): p. 511-22.
69. Ahn, S.H., et al., *Health-related quality of life in disease-free survivors of breast cancer with the general population*. Ann Oncol, 2007. **18**(1): p. 173-82.
70. Nakamura, M., et al., *LVQ-SMOTE - Learning Vector Quantization based Synthetic Minority Over-sampling Technique for biomedical data*. BioData Min, 2013. **6**(1): p. 16.
71. Blagus, R. and L. Lusa, *SMOTE for high-dimensional class-imbalanced data*. BMC Bioinformatics, 2013. **14**: p. 106-106.
72. Song, Y.Y. and Y. Lu, *Decision tree methods: applications for classification and prediction*. Shanghai Arch Psychiatry, 2015. **27**(2): p. 130-5.
73. Wheeler, D.C., et al., *Comparison of ordinal and nominal classification trees to predict ordinal expert-based occupational exposure estimates in a case-control study*. Ann Occup Hyg, 2015. **59**(3): p. 324-35.
74. Upadhyay, S. and N. Patel, *Study of Various Decision Tree Pruning Methods with their Empirical Comparison in WEKA*. International Journal of Computer Applications, 2012. **60**(12): p. 20-25.
75. Dale, M., P. Dale, and P. Tan, *Supervised clustering using decision trees and decision graphs: An ecological comparison*. Ecological modelling, 2007. **204**(1): p. 70-78.
76. Ulla B. Mogensen, H.I., Thomas A. Gerds, *Evaluating Random Forests for Survival Analysis Using Prediction Error Curves*. Journal of Statistical Software, 2012. **50**(11): p. 1-23.
77. Andy Liaw, M.W., *Classification and Regression by randomForest*. R News, 2002. **2/3**: p. 18-22.
78. Shaikhina, T., et al., *Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation*. Biomedical Signal Processing and Control, 2017.
79. Hothorn, T., et al., *Bagging survival trees*. Stat Med, 2004. **23**(1): p. 77-91.

80. Schapire, R.E., *Explaining adaboost*, in *Empirical inference*. 2013, Springer. p. 37-52.
81. Estévez, P.A., et al., *Normalized mutual information feature selection*. IEEE Transactions on Neural Networks, 2009. **20**(2): p. 189-201.
82. Kuncheva, L.I. and S.T. Hadjitodorov. *Using diversity in cluster ensembles*. in *Systems, man and cybernetics, 2004 IEEE international conference on*. 2004. IEEE.
83. Seera, M. and C.P. Lim, *A hybrid intelligent system for medical data classification*. Expert Systems with Applications, 2014. **41**(5): p. 2239-2249.
84. Marshall, R.J., *The use of classification and regression trees in clinical epidemiology*. J Clin Epidemiol, 2001. **54**(6): p. 603-9.
85. Hothorn, T., K. Hornik, and A. Zeileis, *Unbiased recursive partitioning: A conditional inference framework*. Journal of Computational and Graphical statistics, 2006. **15**(3): p. 651-674.
86. Fiteni, F., et al., *Prognostic value of health-related quality of life for overall survival in elderly non-small-cell lung cancer patients*. Eur J Cancer, 2016. **52**: p. 120-8.
87. Mauer, M., et al., *Prognostic factor analysis of health-related quality of life data in cancer: a statistical methodological evaluation*. Expert review of pharmacoeconomics & outcomes research.
88. Eton, D.T., et al., *Early change in patient-reported health during lung cancer chemotherapy predicts clinical outcomes beyond those predicted by baseline report: results from Eastern Cooperative Oncology Group Study 5592*. J Clin Oncol, 2003. **21**(8): p. 1536-43.
89. Braun, D.P., D. Gupta, and E.D. Staren, *Quality of life assessment as a predictor of survival in non-small cell lung cancer*. BMC Cancer, 2011. **11**(1): p. 353.
90. Montazeri, A., *Quality of life data as prognostic indicators of survival in cancer patients: an overview of the literature from 1982 to 2008*. Health and quality of life outcomes, 2009. **7**(1): p. 102.
91. Yun, Y.H., et al., *Prognostic value of quality of life score in disease-free*



- survivors of surgically-treated lung cancer*. BMC Cancer, 2016. **16**: p. 505.
92. Quinten, C., et al., *A global analysis of multitrial data investigating quality of life and symptoms as prognostic factors for survival in different tumor sites*. Cancer, 2014. **120**(2): p. 302-311.
  93. Guinney, J., et al., *Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data*. The Lancet Oncology. **18**(1): p. 132-142.
  94. Roadknight, C., et al. *An ensemble of machine learning and anti-learning methods for predicting tumour patient survival rates*. in *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*. 2015. IEEE.
  95. Hothorn, T., et al., *Survival ensembles*. Biostatistics, 2005. **7**(3): p. 355-373.
  96. Oberije, C., et al., *A Validated Prediction Model for Overall Survival From Stage III Non-Small Cell Lung Cancer: Toward Survival Prediction for Individual Patients*. Int J Radiat Oncol Biol Phys, 2015. **92**(4): p. 935-44.
  97. Zupan, B., et al., *Machine learning for survival analysis: a case study on recurrence of prostate cancer*. Artificial intelligence in medicine, 2000. **20**(1): p. 59-75.
  98. Graf, E., et al., *Assessment and comparison of prognostic classification schemes for survival data*. Statistics in medicine, 1999. **18**(17-18): p. 2529-2545.
  99. Sinisi, S.E., R. Neugebauer, and M.J. van der Laan, *Cross-validated bagged prediction of survival*. Statistical Applications in Genetics and Molecular Biology, 2006. **5**(1).
  100. Montazeri, M., et al., *Machine learning models in breast cancer survival prediction*. Technol Health Care, 2016. **24**(1): p. 31-42.
  101. Langendijk, H., et al., *The prognostic impact of quality of life assessed with the EORTC QLQ-C30 in inoperable non-small cell lung carcinoma treated with radiotherapy*. Radiotherapy and Oncology, 2000. **55**(1): p. 19-25.
  102. Sloan, J.A., et al., *Relationship between deficits in overall quality of life and non-small-cell lung cancer survival*. J Clin Oncol, 2012. **30**(13): p.

- 1498-504.
103. Campbell, P.T., et al., *Associations of recreational physical activity and leisure time spent sitting with colorectal cancer survival*. J Clin Oncol, 2013. **31**(7): p. 876-85.
  104. MO, L., *Physical Activity Guidelines for Americans*. 2008, US Department of Health and Human Services
  105. Maione, P., et al., *Pretreatment quality of life and functional status assessment significantly predict survival of elderly patients with advanced non-small-cell lung cancer receiving chemotherapy: a prognostic analysis of the multicenter Italian lung cancer in the elderly study*. J Clin Oncol, 2005. **23**.
  106. Nowak, A.K., M.R. Stockler, and M.J. Byrne, *Assessing quality of life during chemotherapy for pleural mesothelioma: feasibility, validity, and results of using the European Organization for Research and Treatment of Cancer Core Quality of Life Questionnaire and Lung Cancer Module*. J Clin Oncol, 2004. **22**(15): p. 3172-80.
  107. Langendijk, H., et al., *The prognostic impact of quality of life assessed with the EORTC QLQ-C30 in inoperable non-small cell lung carcinoma treated with radiotherapy*. Radiother Oncol, 2000. **55**.
  108. Ban, W., et al., *Dyspnea as a Prognostic Factor in Patients with Non-Small Cell Lung Cancer*. Yonsei Medical Journal, 2016. **57**(5): p. 1063-1069.
  109. Bottomley, A., et al., *Symptoms and patient-reported well-being: do they predict survival in malignant pleural mesothelioma? A prognostic factor analysis of EORTC-NCIC 08983: randomized phase III study of cisplatin with or without raltitrexed in patients with malignant pleural mesothelioma*. J Clin Oncol, 2007. **25**(36): p. 5770-6.
  110. Nakahara, Y., et al., *Mental state as a possible independent prognostic variable for survival in patients with advanced lung carcinoma*. Cancer, 2002. **94**(11): p. 3006-15.
  111. Wigren, T., *Confirmation of a prognostic index for patients with inoperable non-small cell lung cancer*. Radiother Oncol, 1997. **44**(1): p. 9-15.
  112. Martins, S.J., et al., *Lung cancer symptoms and pulse oximetry in the*

- prognostic assessment of patients with lung cancer*. BMC Cancer, 2005. **5**: p. 72.
113. Sloan, J.A., *Metrics to assess quality of life after management of early-stage lung cancer*. Cancer J, 2011. **17**(1): p. 63-7.
  114. Zhao, X., et al., *Combining gene signatures improves prediction of breast cancer survival*. PLoS One, 2011. **6**(3): p. e17845.
  115. Shin, J., et al., *Combined effect of individual and neighborhood socioeconomic status on mortality in patients with newly diagnosed dyslipidemia: A nationwide Korean cohort study from 2002 to 2013*. Nutr Metab Cardiovasc Dis, 2016. **26**(3): p. 207-15.
  116. Gupta, S., et al., *Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry*. BMJ Open, 2014. **4**(3): p. e004007.
  117. Li, C., et al., *Using the K-nearest neighbor algorithm for the classification of lymph node metastasis in gastric cancer*. Comput Math Methods Med, 2012. **2012**: p. 876545.
  118. Burke, H.B., et al., *Artificial neural networks improve the accuracy of cancer survival prediction*. Cancer, 1997. **79**(4): p. 857-862.
  119. Gao, P., et al., *Which is a more accurate predictor in colorectal survival analysis? Nine data mining algorithms vs. the TNM staging system*. PLoS One, 2012. **7**(7): p. e42015.
  120. Bou-Hamad, I., D. Larocque, and H. Ben-Ameur, *A review of survival trees*. Statistics Surveys, 2011. **5**: p. 44-71.
  121. Ishwaran, H., et al., *Random survival forests for high-dimensional data*. Statistical Analysis and Data Mining: The ASA Data Science Journal, 2011. **4**(1): p. 115-132.
  122. Bühlmann, P. and T. Hothorn, *Boosting algorithms: Regularization, prediction and model fitting*. Statistical Science, 2007: p. 477-505.
  123. Fielding, R. and W.S. Wong, *Quality of life as a predictor of cancer survival among Chinese liver and lung cancer patients*. Eur J Cancer, 2007. **43**.
  124. Brundage, M.D., D. Davies, and W.J. Mackillop, *Prognostic factors in*

- non-small cell lung cancer: a decade of progress.* Chest, 2002. **122**(3): p. 1037-57.
125. Montazeri, A., et al., *Quality of life in lung cancer patients: as an important prognostic factor.* Lung Cancer, 2001. **31**.
  126. Quoix, E., et al., *Carboplatin and weekly paclitaxel doublet chemotherapy compared with monotherapy in elderly patients with advanced non-small-cell lung cancer: IFCT-0501 randomised, phase 3 trial.* Lancet, 2011. **378**(9796): p. 1079-88.
  127. Efficace, F., et al., *Is a patient's self-reported health-related quality of life a prognostic factor for survival in non-small-cell lung cancer patients? A multivariate analysis of prognostic factors of EORTC study 08975.* Ann Oncol, 2006. **17**.
  128. Dharma-Wardene, M., et al., *Baseline FACT-G score is a predictor of survival for advanced lung cancer.* Qual Life Res, 2004. **13**.
  129. Naughton, M., et al., *The health-related quality of life and survival of small-cell lung cancer patients: results of a companion study to CALGB 9033.* Quality of Life Research, 2002. **11**(3): p. 235-248.
  130. Dharma-Wardene, M., et al., *Baseline FACT-G score is a predictor of survival for advanced lung cancer.* Qual Life Res, 2004. **13**(7): p. 1209-16.
  131. Qi, Y., et al., *Pretreatment Quality of Life Is an Independent Prognostic Factor for Overall Survival in Patients with Advanced Stage Non-small Cell Lung Cancer.* Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer, 2009. **4**(9): p. 1075-1082.
  132. Jacot, W., et al., *Quality of life and comorbidity score as prognostic determinants in non-small-cell lung cancer patients.* Ann Oncol, 2008. **19**(8): p. 1458-64.
  133. Sundstrom, S., et al., *Palliative thoracic radiotherapy in locally advanced non-small cell lung cancer: can quality-of-life assessments help in selection of patients for short- or long-course radiotherapy?* J Thorac Oncol, 2006. **1**(8): p. 816-24.
  134. Mountain, C.F., *Revisions in the International System for Staging Lung*

- Cancer*. Chest, 1997. **111**(6): p. 1710-7.
135. Bulzebruck, H., et al., *New aspects in the staging of lung cancer. Prospective validation of the International Union Against Cancer TNM classification*. *Cancer*, 1992. **70**(5): p. 1102-10.
  136. Shah, R., et al., *Results of surgical treatment of stage I and II lung cancer*. *J Cardiovasc Surg (Torino)*, 1996. **37**(2): p. 169-72.
  137. Harpole, D.H., Jr., et al., *A prognostic model of recurrence and death in stage I non-small cell lung cancer utilizing presentation, histopathology, and oncoprotein expression*. *Cancer Res*, 1995. **55**(1): p. 51-6.
  138. Sagawa, M., et al., *Five-year survivors with resected pN2 nonsmall cell lung carcinoma*. *Cancer*, 1999. **85**(4): p. 864-8.
  139. Feinstein, A.R. and C.K. Wells, *A clinical-severity staging system for patients with lung cancer*. *Medicine (Baltimore)*, 1990. **69**(1): p. 1-33.
  140. Albain, K.S., et al., *Survival determinants in extensive-stage non-small-cell lung cancer: the Southwest Oncology Group experience*. *J Clin Oncol*, 1991. **9**(9): p. 1618-26.
  141. Takigawa, N., et al., *Prognostic factors for patients with advanced non-small cell lung cancer: univariate and multivariate analyses including recursive partitioning and amalgamation*. *Lung Cancer*, 1996. **15**(1): p. 67-77.
  142. Ganz, P.A., J.J. Lee, and J. Siau, *Quality of life assessment. An independent prognostic variable for survival in lung cancer*. *Cancer*, 1991. **67**(12): p. 3131-5.
  143. Coates, A., F. Porzsolt, and D. Osoba, *Quality of life in oncology practice: Prognostic value of EORTC QLQ-C30 scores in patients with advanced malignancy*. *European Journal of Cancer*. **33**(7): p. 1025-1030.
  144. Herndon, J.E., 2nd, et al., *Is quality of life predictive of the survival of patients with advanced nonsmall cell lung carcinoma?* *Cancer*, 1999. **85**(2): p. 333-40.
  145. Buccheri, G., *Depressive reactions to lung cancer are common and often followed by a poor outcome*. *Eur Respir J*, 1998. **11**(1): p. 173-8.
  146. Burrows, C.M., W.C. Mathews, and H.G. Colt, *Predicting survival in*

*patients with recurrent symptomatic malignant pleural effusions: an assessment of the prognostic values of physiologic, morphologic, and quality of life measures of extent of disease.* Chest, 2000. **117**(1): p. 73-8.

## 국문 초록

**서론:** 폐암 생존 예측은 성공적인 폐암생존자들의 암 생존 이후의 계획에 중요한 요소로 알려져 있다. 본 연구의 목적은 인구-사회-임상 및 HRQOL 변수를 이용하여 무병 폐암 생존자의 5 년 생존 예측 모델을 구축하고 기존의 알려진 임상 변수를 기반으로 예측 모델과 예측 모델을 비교하는 것이다. 이에, Cox 비례 위험 모델 및 기계 학습 기술 (Machine Learning Techniques)과 같은 다양한 알고리즘을 생존예측 모델링 프로세스에 적용해보았다.

**방법:** 본 연구는 1994 년부터 2002 년까지 2 개의 국내 대학 병원에서 폐암 수술을 받은 809 명의 환자에 관한 자료를 바탕으로 진행되었으며, 연령, 성별, 병기, 학력 및 소득을 포함한 임상 및 사회 인구 통계 학적 변수와 EORTC QLQ-C30 설문을 통한 건강 관련 삶의 질, 건강 습관 요소 등의 환자 보고 성과 (Patient Reported Outcome) 지표 중, 체계적인 문헌 고찰 및 단변량 생존 분석을 사용하여 예후 모형에 사용될 독립 변수를 선택하였다. 건강 관련 삶의 질 점수 평가를 위해서 EORTC QLQ-C30, 폐암 생존자 특이적 건강관련 삶의 질 설문인 QLQ-LC13, 병원 불안 및 우울증 척도 (HADS), 및 외상 후 긍정적 성장 (PTGI) 등이 평가되었으며. 생존자의 수술 전 BMI 와 신체 활동도 예후 인자로 선택 되었다. WHO 가 제시한 ICF 분류 기준에 따라, 3 종류의 Feature set 을 구성하였으며, 세 가지 예측 모델링 Feature set 은 다음과 같다. 1) 임상 및 사회 - 인구 통계 학적 변수, 2) HRQOL 및 라이프 스타일 요인 만 고려한 feature set, 3) Feature set 1, 2 의 변수들이 모두 고려된 Feature set 이다. 먼저, 각각의 Feature set 에 대해 C-statistic 및 Hosmer-Lemeshow 카이 제곱 통계를 사용하여 3 가지 Cox 비례 위험 회귀 모델을 구성하고 개별 모형의 성능을 비교했다. 그 다음으로, 의사 결정 트리 (DT), 랜덤

포리스트 (RF), 배깅 및 적응 형 부스트 (AdaBoost) 를 사용하는 4 가지 기계 학습 알고리즘을 세 가지 Feature set 에 적용하고 서로의 성능과 정확도 값을 비교하였다. MLT 를 기반으로 한 파생 된 예측 모델의 성능은 K- fold 교차 유효성 검증에 의해 내부적으로 검증되었다.

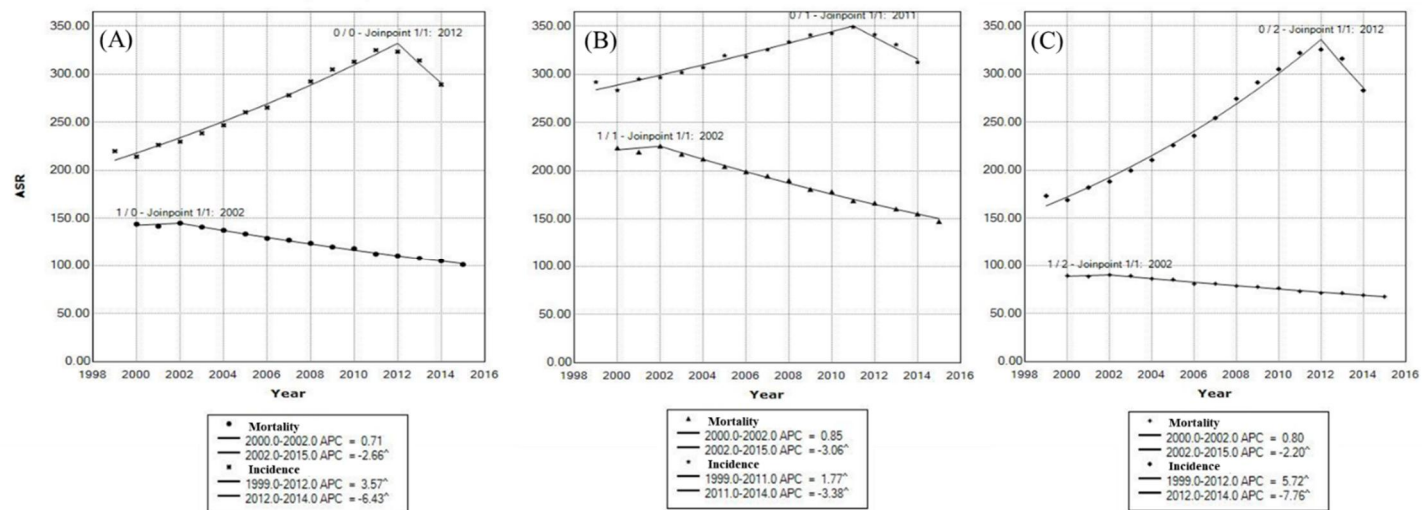
**결과:** Cox 모델링에서 모델 Cox-3 (Feature set 3: HRQOL 요소가 임상 및 사회 인구 통계 학적 변수에 추가됨) 모델은 두 개의 다른 Cox 회귀 분석 (Cox-1, 2 모델)에 비해 AUC 값이 더 높았다. MLT 기반으로 모델링 방법을 적용했을 때 가장 효과적인 모델은 DT 의 DT-3 모델, RF 의 모델 RF-3, Bagging 의 Model Bag-3, AdaBoost 의 모델 AdaBoost-3 이었으며, 각 MLT 의 정확도는 Model RF-3, Model Bag-3 모델 AdaBoost-3 모델에서 가장 높은 정확도 및 모델 AUC 를 나타냈다.

**결론:** 임상 및 사회 인구 통계적 변수에 HRQOL 및 건강습관 요인이 추가된 모델이 Cox 비례 위험 회귀 모형에서 가장 성능이 우수했으며, 이 패턴은 MLT 알고리즘을 적용한 폐암 생존 예측모형에서도 적용되는 것을 확인 할 수 있었다. HRQOL 및 건강습관 요인이 추가된 모델을 통해, 폐암 생존 예측 모델의 정확도가 향상되면 의사와 생존자가 향후 계획 및 암 치료에 대한 지원에 대해보다 의미 있는 결정을 내릴 수 있을 것이다.



# **APPENDIX**

**Supplementary figures and tables of  
lung cancer survival prognostic factor prediction models**



Appendix Figure 1. Trends in age standardized rates in all cancer incidents and mortality in Korea based on the joint point regression  
(A) Both sexes (B) Male (C) Female

Appendix Table 1. Systematic Review Search Term Steps and Each of Counts

Search Term	Counts
1      “disease-free” OR ("lung neoplasms"[MeSH Terms] OR ("lung" AND "neoplasms") OR "lung neoplasms" OR ("lung" AND "cancer" ) OR "lung cancer" ) AND ("survivors"[MeSH Terms] OR "survivors" OR "survivor")	4,594
2      "prognostic factors" OR "predictor" OR "Predictive" OR "determinant" OR "Predict" OR "survival"	2,901
3      "HRQOL" OR "health-related quality of life OR quality of life" OR "patient-reported outcomes" OR "BMI"OR "Physical Activity"OR "depression"OR "anxiety" OR "Post-traumatic growth" OR "PTGI"	460

Appendix Table 2. Prognostic factors of age

No	First author	Year	No. of patients	Target cancer type	Study design	Prognosis factor	Categories	aHR	Primary endpoint	Quality assessment	Ref
1	Young Ho Yun et al.	2016	809	Disease-free lung cancer survivor	Cross-sectional study	Age, years	< 65 versus $\geq$ 65	1.84	OS	1	[91]
2	Wooho Ban et al.	2016	457	NSCLC	Cross-sectional study	Age, years	$\leq$ 68 versus >68	1.60	OS	1	[108]
3	Sloan JA et al.	2012	2,442	NSCLC	Cross-sectional study	Age, years	> 80 versus 50 to 65	0.36	OS	1	[102]
							> 80 versus 65 to 80	0.56			
							> 80 versus $\leq$ 50	0.35			
4	Richard Fielding et al.	2007	358	NSCLC	Longitudinal study	Age, years	Younger versus older age	1.02	OS	3	[123]
5	Michael D. Brundage et al.	2002	1,960	NSCLC	Meta analysis prognostic factor	Age, years	NA	NA	OS	1	[124]
6	Montazeri et al.	2001	129	NSCLC SCLC	Cross-sectional	Age, years	Continuous	1.10	OS	1	[125]

Appendix Table 3. Prognostic factors of sex

No	First author	Year	No. of patients	Target cancer Type	Study design	Prognosis factor	Categories	aHR	Primary endpoint	Quality assessment	Ref
1	YH Yun et al.	2016	809	Disease-free lung cancer survivor	Cross-sectional Study	Sex	Women versus men	2.52	OS	1	[91]
2	Braun et al.	2011	1,194	NSCLC	cross-Sectional study	Sex	Male versus female	0.78	OS	1	[89]
3	Quoix E et al.	2011	451	NSCLC	RCT	Sex	Male versus female	0.77	OS	1	[126]
4	Efficace F et al.	2006	391	Advanced NSCLC	cross-Sectional study	Sex	Female versus male	1.32	OS	1	[127]
5	Maione et al.	2005	566	Advanced NSCLC	RCT	Sex	Male versus female	0.78	OS	1	[105]
6	Dharma-Wardene et al.	2004	44	Advanced lung cancer	Cross-sectional study	Sex	Male versus female	0.32	OS	1	[128]
7	Nowak et al.	2004	53	Pleural mesothelomia	RCT	Sex			OS	1	[106]
8	Nakahara et al.	2002	179	Advanced NSCLC	Cross-sectional study	Sex	Female versus male	2.35	OS	3	[110]
9	Naughton MJ et al.,	2002	70	SCLC	Cohort	Sex	Male versus female	0.488	OS	1	[129]
10	Michael D. Brundage et al.	2002	1,960	NSCLC	Systematic-review	Sex			OS	3	[124]

Appendix Table 4. Prognostic factors of stage

No	First author	Year	No. of patients	Target cancer type	Study design	Prognosis factor	Categories	aHR	Primary endpoint	Quality assessment	Ref
1	YH Yun et al.	2014	809	Disease-free lung cancer survivor	Cross-sectional study	Stage	0–I versus II–III	1.7	OS	1	[91]
2	Wooho Ban et al.	2016	457	NSCLC	Cross-sectional study	Stage	0–I versus II–III	1.6	OS	1	[108]
3	Braun et al.	2011	1,194	NSCLC	Cross-sectional study	Stage at diagnosis	Locoregional disease as reference	1.67	OS	1	[89]
4	Richard Fielding et al.	2007	358	NSCLC	Cross-sectional study	Cancer stage	Advanced versus less advanced	1.978	OS	1	[123]
							IV versus III	0.59	OS	1	
5	Sloan JA et al.	2012	2,442	NSCLC	Cross-sectional study	Stage	IV versus II	0.67	OS	1	[102]
							IV versus I	0.39	OS		
6	Nakahara et al.	2002	179	Advanced NSCLC	Cross-sectional	Stage	I versus IV	1.59	OS	1	[110]
7	Dharma-Wardene et al.	2004	44	Advanced lung cancer	Cross-sectional	Stage	III versus IV	0.94	OS	1	[130]
8	Martins SJ et al.	2005	41	locally advanced or metastatic lung cancer	Longitudinal study	Clinical stage	III versus IV	2.18	OS	3	[112]
9	Bottomley et al.	2007	250	malignant pleural mesothelioma	Cross-sectional study	Stage		1.396	OS	1	[109]

Appendix Table 5. Prognostic factors of treatment type

No	First author	Year	No. of patients	Target cancer type	Study design	Prognosis factor	Categories	aHR	Primary endpoint	Quality assessment	Ref
1	Braun et al.	2011	1,194	NSCLC	Cross-sectional study	Prior treatment history	Previously treated as reference	0.55	OS	1	[89]
2	Fiteni F et al.,	2016	451	NSCLC patients aged 70-89	Cross-sectional study	Treatment type	Mono chemotherapy versus Doublet chemotherapy	0.64	OS	1	[86]
3	Sloan JA et al.	2012	2,442	NSCLC	Cross-sectional study	Treatment	Radiation plus chemotherapy	1.55	OS	1	[102]
							Surgery	0.36	OS	1	
4	Quoix E et al.	2011	451	NSCLC	Longitudinal study	Treatment	Mono chemotherapy versus Doublet chemotherapy	0.64	OS	3	[126]

Appendix Table 6. Prognostic factors of other clinical factors (meta, recurrence, comorbidity, and FEV1/FVC)

No	First author	Year	No. of patients	Target cancer type	Study design	Prognosis factor	Categories	aHR	Primary endpoint	Quality assessment	Ref
1	Fiteni F et al.,	2016	451	NSCLC patients aged 70-89	Cross-sectional	Brain metastases	Mono chemotherapy versus Doublet chemotherapy	0.64	OS	1	[86]
2	Dharma-Wardene et al.	2004	44	Advanced lung cancer	Cross-sectional	Hepatic metastases	Total Score Absent versus present	0.94 0.47	OS	0 0	[130]
3	Montazeri et al.	2001	129	NSCLC and SCLC	Cross-sectional	Extent of disease		1.1	OS	1	[125]
4	Maione et al.	2005	566	Advanced NSCLC	RCT	No. of sites of disease	For each added site	0.78	OS	1	[105]
5	Wooho Ban et al.	2016	457	NSCLC	cross-sectional study	FEV1/FVC, % FEV1	≥70 versus <70 % predicted	1 0.99	OS OS	1 1	[108]
6	Michael D. Brundage et al.	2002	1,960	NSCLC	Systematic-review	“T” factor ”N” factor	[3] [5-8]	NA	OS	3	[124]
7	Sloan JA et al.	2012	2,442	NSCLC		Recurrence and/or progression	Yes versus no	0.51	OS	1	[102]



Appendix Table 7. Prognostic factors of HRQOL (Physical functioning)

No	First author	Year	No. of patients	Target cancer type	Study design	Prognosis factor	Categories	aHR	Primary endpoint	Quality assessment	Ref
1	YH Yun et al.	2014	809	Disease-free lung cancer survivor	Cross-sectional study	Physical functioning	> 33.33 versus ≤ 33.33	2.39	OS	1	[91]
2	Benjamin Movsas et al.	2009	239	NSCLC	Cross-sectional study	Physical functioning	< 66.66 versus ≥ 66.66	1.69	OS	1	[38]
3	Braun et al.	2011	1,194	NSCLC	Cross-sectional study	Physical functioning		0.99	OS	1	[89]
4	Fiteni F et al.,	2016	451	NSCLC patients aged 70-89	Cross-sectional study	Physical functioning		0.98	OS	1	[86]
5	Nowak et al.	2004	53	Pleural mesothelioma	Longitudinal study	Physical function (QLQ-C30)		No HR suggested.	OS	3	[106]
6	SteinSundstrøm et al.	2006	301	Stage III non-small-cell lung cancer	Cross-sectional study	Physical function	< median versus ≥ median	1.5	OS	1	[133]

Appendix Table 8. Prognostic factors of HRQOL (Global QOL)

No	First author	Year	No. of patients	Target Cancer Type	study design	Prognosis Factor	Categories	aHR	Primary Endpoint	Quality assessment	Ref
1	Yingwei Qi et al.,	2010	355	advanced NSCLC	cross-sectional study	QOL uniscale	low vs. high scores	1.63	OS	1	[131]
2	Michael D. Brundage et al.,	2002	1,960	NSCLC	Systematic-review	Quality of life	[12-14]	Na	OS	2	[124]
3	Benjamin Movsas et al.,	2009	239	NSCLC	cross-sectional study	global QOL score	<66.66 vs. ≥66.66	1.69	OS	1	[38]
4	Braun et al	2011	1,194	NSCLC	cross-sectional study	Global QOL		0.001	OS	1	[89]
5	Langendijk H et al.,	2000	NA	NSCLC	cross-sectional study	Global QoL	EORTC QLQ-C30 Lung Cancer Symptom Scale (>50 vs. <50)	0.993	OS	1	[107]
6	Fiteni F et al.,	2016	451	NSCLC patients aged 70-89	Cross-sectional	Global health status		0.98	OS	1	[86]
7	Sloan JA et al.,	2012	2,442	NSCLC	Cross-sectional	QOL		1.55	OS	1	[102]
8	Jacot W et al.,	2008	301	NSCLC	Cross-sectional	QoL	≤22.2 vs. >22.2	2.2	OS	1	[132]
9	Montazeri et al.	2001	129	NSCLC and SCLC	Cross-sectional	Global quality of life	High, vs. low	3.2	OS	1	[125]
10	Dharma-Wardene et al	2004	44	Advanced lung cancer	Cross-sectional	Baseline FACT-G	Total Score	0.94	OS	1	[128]
11	Maione et al	2005	566	Advanced NSCLC	RCT	Quality of Life	Better vs. intermediate Better vs. worse	1.62 1.76	OS OS	2 2	[105]
12	Brown et al	2005	273	NSCLC	RCT	Global QOL		Na	OS	2	
13	Bottomley et al	2007	250	malignant pleural mesothelioma	Cross-sectional	Global health status/QOL		0.868	OS	1	[109]

Appendix Table 9. Prognostic factors of HRQOL (other functioning)

No	First author	Year	No. of patients	Target Cancer Type	study design	Prognosis Factor	Categories	aHR	Primary Endpoint	Quality assessment	Ref
1	Braun et al	2011	1,194	NSCLC	cross-sectional study	Emotional	NA	1.003	OS	2	[89]
2	Langendijk H et al.,	2000		NSCLC	cross-sectional study	Role functioning	NA	0.996	OS	1	[107]
3	Fiteni F et al.,	2016	451	NSCLC patients aged 70-89	Cross-sectional	Role functioning	NA	0.99	OS	1	[86]
4	Nowak et al	2004	53	Pleural mesothelioma	RCT	Emotional function	NA	Only p-value were suggested	OS	2	[106]
5	Brown et al	2005	273	NSCLC	RCT	role functioning	NA		OS	2	
6	SteinSundstrøm et al.,	2006	301	stage III non-small-cell lung cancer	Cross-sectional data	Role function	< median vs. ≥ median	0.63	OS	1	[133]
7	Bottomley et al	2007	250	malignant pleural mesothelioma	Cross-sectional data	Cognitive functioning	NA	0.892	OS	1	[109]
						Social functioning	NA	0.916	OS	1	

Appendix Table 10. Prognostic factors of HRQOL (Dyspnea)

No	First author	Year	No. of patients	Target Cancer Type	study design	Prognosis Factor	Categories	aHR	Primary Endpoint	Quality assessment	Ref
1	YH Yun et al.,	2014	809	disease-free lung cancer survivor	cross-sectional study	dyspnea	<66.66 vs. ≥66.66	1.56	OS	1	[91]
2	Benjamin Movsas et al.,	2009	239	NSCLC	cross-sectional study	Dyspnea (LC13)	<86.7 vs. ≥86.7	NA	OS	1	[38]
3	Wooho Ban et al.,	2016	457	NSCLC	cross-sectional study	Dyspnea	(-) vs. (+) mMRC 0-1	1.15	OS	1	[108]
							(-) vs. (+) mMRC ≥2	1.84	OS	1	
4	Langendijk H et al.,	2000		NSCLC	cross-sectional study	Dyspnea		1.005	OS	1	[107]
5	Bottomley et al	2007	250	malignant pleural mesothelioma		Dyspnea		1.106	OS	1	[109]

Appendix Table 11. Prognostic factors of HRQOL (Appetite Loss)

No	First author	Year	No. of patients	Target Cancer Type	study design	Prognosis Factor	Categories	aHR	Primary Endpoint	Quality assessment	Ref
1	Braun et al	2011	1,194	NSCLC	cross-sectional study	Appetite Loss		1.003	OS	1	[89]
2	Langendijk H et al.,	2000		NSCLC	cross-sectional study	Appetite loss		1.007	OS	1	[107]
3	Fiteni F et al.,	2016	451	NSCLC patients aged 70-89	cross-sectional study	Appetite loss		1.007	OS	1	[109]
4	Richard Fielding et al.,	2007	358	NSCLC		Eating appetite		0.888	OS	1	[123]
5	Brown et al	2005	273	NSCLC	RCT	Appetite loss		NA	OS	2	
6	Martins SJ et al	2005	41	locally advanced or metastatic lung cancer	RCT	Appetite loss		1.14	OS	2	[112]
7	SteinSundstrøm et al.,	2006	301	stage III non-small-cell lung cancer	cross-sectional study	Appetite loss	> median vs. ≤ median	1.84	OS	1	[133]
8	Bottomley et al	2007	250	malignant pleural mesothelioma		Appetite loss		1.083	OS	1	[109]

Appendix Table 12. Prognostic factors of HRQOL (Fatigue)

No	First author	Year	No. of patients	Target Cancer Type	study design	Prognosis Factor	Categories	aHR	Primary Endpoint	Quality assessment	Ref
1	Fiteni F et al.,	2016	451	NSCLC patients aged 70-89	cross-sectional study	Fatigue (QLQ-C30)		1.011	OS	1	[86]
2	Nowak et al	2004	53	Pleural mesothelioma	RCT	Fatigue (QLQ-C30)		NA	OS	2	[106]
3	Brown et al	2005	273	NSCLC	RCT	Fatigue (QLQ-C30)		NA	OS	2	
4	Martins SJ et al	2005	41	locally advanced or metastatic lung cancer	RCT	Fatigue (QLQ-C30)		1.23	OS	2	[112]
5	Bottomley et al	2007	250	malignant pleural mesothelioma	cross-sectional study	Fatigue (QLQ-C30)		1.162	OS	1	[109]

Appendix Table 13. Prognostic factors of HRQOL (Pain)

No	First author	Year	No. of patients	Target Cancer Type	study design	Prognosis Factor	Categories	aHR	Primary Endpoint	Quality assessment	Ref
1	Fiteni F et al.,	2016	451	NSCLC patients aged 70-89	cross-sectional study	Pain		1.007	OS	1	[86]
2	Efficace F et al.,	2006	391	advanced NSCLC	cross-sectional study	pain		1.11	OS	1	[127]
3	Nowak et al	2004	53	Pleural mesothelioma	RCT	Composite pain score (LC13)		NA	OS	2	[106]
4	Bottomley et al	2007	250	malignant pleural mesothelioma	cross-sectional study	LC pain chest		1.092	OS	1	[109]

Appendix Table 14. Prognostic factors of HRQOL (other symptoms)

No	First author	Year	No. of patients	Target Cancer Type	study design	Prognosis Factor	Categories	aHR	Primary Endpoint	Quality assessment	Ref
1	Michael D. Brundage et al,	2002	1,960	NSCLC	systematic-review	Symptoms	[9, 15]	NA	OS	2	[124]
2	Sloan JA et al.,	2012	2,442	NSCLC	cross-sectional study	Lung Cancer Symptom Scale (>50 vs. <50)		1.55	OS	1	[102]
3	Fiteni F et al.,	2016	451	NSCLC patients aged 70-89	cross-sectional study	Nausea		1.007	OS	1	[86]
4	Efficace F et al.,	2006	391	advanced NSCLC	cross-sectional study	Dysphagia		1.04	OS	1	[127]
5	Brown et al	2005	273	NSCLC	RCT	constipation		NA	OS	2	
						Nausea and vomiting		1.158	OS	1	[37]
6	Bottomley et al	2007	250	malignant pleural mesothelioma	cross-sectional study	LC coughing		1.003	OS	1	
						LC dysphagia		1.134	OS	1	[109]
						LC peripheral neuropathy		0.898	OS	1	



Appendix Table 15. Prognostic factors of HRQOL (Performance Score)

No	First author	Year	No. of patients	Target Cancer Type	study design	Prognosis Factor	Categories	aHR	Primary Endpoint	Quality assessment	Ref
1	Yingwei Qi et al.,	2010	355	advanced NSCLC	cross-sectional study	Performance Score (PS)	1-2 vs. 0	1.22	OS	1	[131]
2	Michael D. Brundage et al.,	2002	1,960	NSCLC	Systematic-review	PS	[10, 11]		OS	2	[124]
3	Benjamin Movsas et al.,	2009	239	NSCLC	cross-sectional study	Karnofsky performance score	70-80 vs. 90-100	1.47	OS	1	[38]
4	Wooho Ban et al.,	2016	457	NSCLC	cross-sectional study	PS (ECOG)	0-2 vs. $\geq 3$	3.67	OS	1	[108]
5	Fiteni F et al.,	2016	451	NSCLC patients aged 70-89	cross-sectional study	Performance status	0-1 vs. 2	2.10	OS	1	[86]
6	Efficace F et al.,	2006	391	advanced NSCLC	cross-sectional study	Performance status	0-1 vs. 2	1.63	OS	1	[127]
7	Sloan JA et al.,	2012	2,442	NSCLC	cross-sectional study	ECOG performance status	2,3,4 vs. 1	0.53	OS	1	[102]
8	Quoix E et al.,	2011	451	NSCLC	RCT	Performance status	0-1 vs. 2		OS	2	[126]
9	Burrows CM et al.,	2000	85	recurrent symptomatic malignant pleural effusions		Karnofsky Performance Scale	$<70$ vs $\geq 70$	0.73	OS	1	[146]
10	Nakahara et al	2002	179	Advanced NSCLC	cross-sectional study	ECOG PS		1.87	OS	1	[110]
11	Dharma-Wardene et al	2004	44	Advanced lung cancer	cross-sectional study	ECOG PSR	0 vs. 1 0 vs. 2	0.49 0.24		1	[128]
12	Nowak et al	2004	53	Pleural mesothelioma	RCT	ECOG performance status 2				2	[106]
13	Maione et al	2005	566	Advanced NSCLC	RCT	Performance status	0-1 vs. 2	1.46		2	[105]

Appendix Table 16. Prognostic factors of HRQOL (BMI and weight loss)

No	First author	Year	No. of patients	Target Cancer Type	study design	Prognosis Factor	Categories	aHR	Primary Endpoint	Quality assessment	Ref
1	YH Yun et al.,	2014	809	disease-free lung cancer survivor	cross-sectional study	BMI	$\geq 23$ vs. $< 23$	1.75	OS	1	[91]
2	Yingwei Qi et al.,	2010	355	advanced NSCLC	cross-sectional study	BMI	Underweight vs. normal	1.87		1	
							Overweight vs. normal	0.86	OS	1	[131]
							Obese vs. normal	0.82		1	
3	Michael D. Brundage et al.,	2002	1,960	NSCLC	Systematic-review	Weight loss	[29]		OS	1	[124]
4	Benjamin Movsas et al.,	2009	239	NSCLC	cross-sectional study	Global QOL score	$< 66.66$ vs. $\geq 66.66$	1.69	OS	1	[38]
12	Eton DT et al.,	2003	573	NSCLC	cross-sectional study	PWB score		1.09	OS	1	[88]
										2	
							$\leq 20$ vs. $> 20 - \leq 26$	0.87			
16	Quoix E et al.,	2011	451	NSCLC	RCT	BMI ( $\text{kg}/\text{m}^2$ )	$\leq 20$ vs. $> 26 - \leq 30$	0.74	OS	1	[126]
							$\leq 20$ vs. $> 30$	0.78			
						Weight loss(%)	$> 5$ vs. $\leq 5$	0.56		1	
19	Nakahara et al	2002	179	Advanced NSCLC	Cross-sectional	Weight loss $> 5\%$		1.48	OS	1	[110]

Appendix Table 17. Comparisons of EORTC QLQ-C30 scores between event group and no-event groups

Variable	No event (n=713)		Event (n=96)		p-value
	Mean	SD	Mean	SD	
EORTC QLQ-C30					
Physical functioning	76.42	18.98	67.08	24.83	0.001
Role functioning	81.51	23.22	71.70	27.90	0.001
Emotional functioning	83.72	19.42	77.95	22.62	0.008
Cognitive functioning	80.93	19.77	75.17	21.63	0.015
Social functioning	85.11	23.02	77.78	26.34	0.004
Global QOL	61.63	19.84	56.42	19.96	0.018
Fatigue	29.31	23.90	37.38	23.76	0.002
Nausea/Vomiting	6.19	14.92	5.56	14.03	0.692
Pain	17.06	22.58	22.40	24.99	0.032
Dyspnea	32.72	30.67	46.18	33.98	<0.001
Insomnia	19.12	27.94	21.87	30.52	0.404
Appetite loss	15.52	25.64	26.74	32.67	<0.001
Constipation	11.97	22.76	14.93	26.87	0.242
Diarrhea	8.04	18.30	11.11	23.03	0.136
Financial difficulties	18.75	28.69	28.82	33.02	<0.001

Appendix Table 18. Comparisons of EORTC QLQ-LC13 scores between event group and no-event groups

group and no event groups					
Variable	No event (n=713)		Event (n=96)		p-value
	Mean	SD	Mean	SD	
EORTC QLQ-LC13					
Dyspnea	25.77	22.02	40.16	27.33	<0.001
Coughing	16.74	22.84	27.78	30.46	<0.001
Hemoptysis	1.50	8.34	2.08	9.44	0.563
Sore mouth	6.88	16.80	10.07	20.59	0.09
Dysphagia	5.00	15.95	8.33	21.08	0.149
Peripheral neuropathy	15.10	24.75	17.36	27.35	0.407
Alopecia	11.10	22.23	12.85	23.88	0.473
Pain in chest	17.67	24.78	21.53	29.41	0.222
Pain in arm or shoulder	20.48	27.51	20.83	28.72	0.906
Pain in other parts	11.83	24.07	14.03	26.44	0.407

Appendix Table 19. Comparisons of PTGI and HADS scores between event group and no-event groups

		No event (n=713)		Event (n=96)		p-value
Variable		n	(%)	n=131	7.1%	
PTGI						
Relating to others (35)	≥23	297	0.895	416	0.872	0.331
	<23	35	0.105	61	0.128	
New possibilities (25)	≥18	164	92.7%	501	86.7%	0.032
	<18	13	7.3%	77	13.3%	
Personal strength (20)	≥15	223	94.1%	490	85.7%	0.001
	<15	14	5.9%	82	14.3%	
Spiritual change (10)	≥5	367	90.2%	346	86.1%	0.071
	<5	40	9.8%	56	13.9%	
Appreciation of life (15)	≥11	328	90.9%	385	85.9%	0.031
	<11	33	9.1%	63	14.1%	
HADS						
Anxiety	<8	575	90.7%	134	78.8%	p <0.001
	≥8	59	9.3%	36	21.2%	
Depression	<8	445	90.8%	262	83.7%	0.002
	≥8	45	9.2%	51	16.3%	

Appendix Table 20. Comparisons of socio-demographic and clinical characteristics of training and test set based on SMOTE data

Variable		Balanced data with SMOTE *				
		Training dataset (n=735)		Testing dataset (n=317)		P- value
		N	%	N	%	
Death	No	406	0.705	170	0.295	0.695
	Yes	333	0.694	147	0.306	
Age, years		739		317		0.243
	< 65	358	68.3%	166	31.7%	
	≥ 65	381	71.6%	151	28.4%	
Sex	Female	600	70.4%	252	29.6%	0.522
	Male	139	68.1%	65	31.9%	
Monthly income (USD)				1		0.597
	≥ 3,000	173	68.7%	79	31.3%	
	< 3,000	566	70.4%	238	29.6%	
Education	≥ High school degree	265	71.0%	108	29.0%	0.577
	< High school degree	474	69.4%	209	30.6%	
Employment status	Yes	273	68.3	127	31.8	0.338
	No	466	71.0	190	29.0	
Currently married	Yes	659	70.1%	281	29.9%	0.800
	No	80	69.0%	36	31.0%	
FEV1/FVC	(FEV1/FVC)*100≥0.7	332	68.6%	152	31.4%	0.484
	(FEV1/FVC)*100<0.7	382	70.6%	159	29.4%	
Stage	Stage 0— I	425	71.9%	166	28.1%	0.123
	Stage II—III	314	67.5%	151	32.5%	
Recurrence	No	509	69.2%	227	30.8%	0.376
	Yes	230	71.9%	90	28.1%	
Local invasion of tumor	No	210	68.9%	95	31.1%	0.595
	Yes	526	70.5%	220	29.5%	
Lymph node metastasis	No	488	70.5%	204	29.5%	0.565
	Yes	247	68.8%	112	31.2%	
Treatment type	OP	402	70.0%	172	30.0%	0.357
	OP+RT	65	63.1%	38	36.9%	
	OP+CT	159	72.9%	59	27.1%	
	OP+CT+RT	109	69.4%	48	30.6%	

Appendix Table 21. Comparisons of EORTC-QLQ-C30 variables on training and test set based on SMOTE data

Test Set based on SMOTE *					
Variable	Balanced data with SMOTE *				p-value
	Training dataset (n=735)		Testing dataset (n=317)		
	N	%	N	%	
EORTC QLQ-C30					
Physical functioning	69.0	21.1	69.8	22.0	0.584
Role functioning	73.6	26.7	74.9	27.1	0.477
Emotional functioning	83.1	17.5	82.2	19.3	0.471
Cognitive functioning	82.0	17.7	81.4	18.9	0.583
Social functioning	86.5	79.8	86.8	21.6	0.848
Global QOL	61.9	17.9	62.2	18.4	0.820
Fatigue	32.8	19.7	32.4	21.0	0.753
Nausea/Vomiting	5.5	12.8	601	14.2	0.518
Pain	17.8	20.8	17.4	20.8	0.802
Dyspnea	41.7	29.8	40.3	29.5	0.505
Insomnia	17.1	25.7	18.0	24.7	0.615
Appetite loss	19.6	27.4	20.5	29.5	0.617
Constipation	10.3	20.4	901	18.9	0.350
Diarrhea	6.7	16.1	6.8	16.6	0.935
Financial difficulties	20.3	25.9	19.9	25.8	0.804

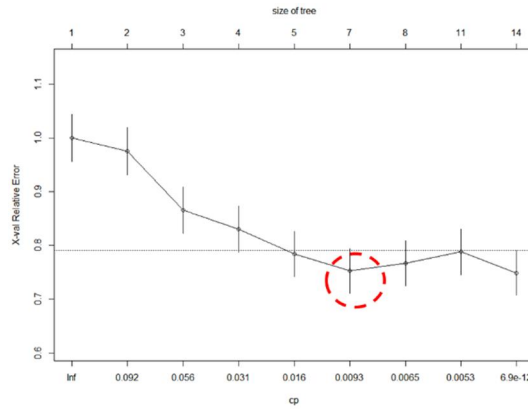
Appendix Table 22. Comparisons of EORTC-QLQ-LC13 variables on training and test set based on SMOTE data

		Balanced data with SMOTE *				
		Training dataset (n=735)		Testing dataset (n=317)		p-value
		N	%	N	%	
EORTC QLQ-LC13						
Dyspnea	33.9	23.8	32.8	24.9	0.518	
Coughing	19.0	22.2	20.3	23.0	0.371	
Hemoptysis	2.9	9.2	2.7	8.1	0.717	
Sore mouth	10.6	15.7	11.7	17.1	0.331	
Dysphagia	6.3	16.4	6.5	16.1	0.906	
Peripheral neuropathy	15.9	22.8	13.8	22.1	0.152	
Alopecia	10.1	18.6	11.7	20.1	0.191	
Pain in chest	20.8	26.9	22.8	27.9	0.278	
Pain in arm or shoulder	21.7	27.5	21.0	28.8	0.720	
Pain in other parts	10.9	22.0	10.4	22.4	0.712	



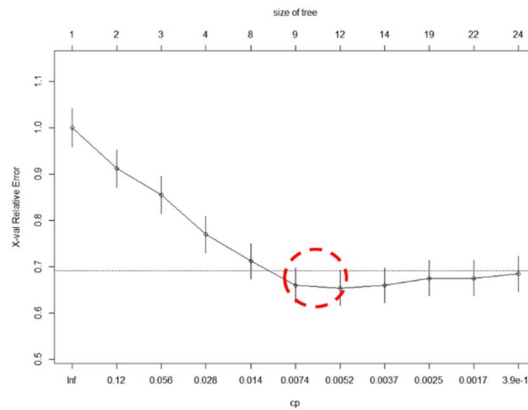
Appendix Table 23. Comparisons of EORTC-QLQ-LC13 variables on training and test set based on SMOTE data

	Balanced data with SMOTE *				
	Training dataset (n=735)		Testing dataset (n=317)		p-value
	N	%	N	%	
PTGI					
Relating to others (35)	19.9	6.8	20.4	6.9	0.370
New Possibilities (25)	12.7	5.4	13.2	5.3	0.196
Personal Strength (20)	10.9	4.2	11.2	4.2	0.373
Spiritual Change (10)	4.1	2.7	4.1	2.7	0.691
Appreciation of life (15)	8.9	3.2	9.2	3.1	0.269
HADS					
Anxiety	4.2	3.3	4.4	3.8	0.334
Depression	6.2	3.5	6.1	3.4	0.849



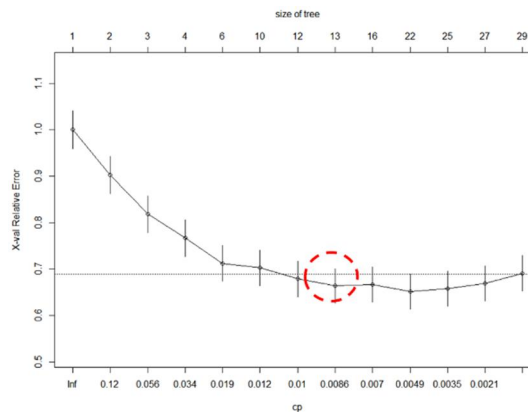
	CP	nsplit	rel error	xerror	xstd
1	1.2721e-01	0	1.00000	1.00000	0.043676
2	6.7138e-02	1	0.87279	0.97527	0.043584
3	4.5936e-02	2	0.80565	0.86572	0.042900
4	2.1201e-02	3	0.75972	0.83039	0.042580
5	1.2367e-02	4	0.73852	0.78445	0.042087
6	7.0671e-03	6	0.71378	0.75265	0.041695
7	5.8893e-03	7	0.70671	0.76678	0.041874
8	4.7114e-03	10	0.68905	0.78799	0.042128
9	1.0000e-20	13	0.67491	0.74912	0.041648

Appendix Figure 2. CP plot and table for model DT-1



	CP	nsplit	rel error	xerror	xstd
1	1.6970e-01	0	1.00000	1.00000	0.040953
2	7.8788e-02	1	0.83030	0.91212	0.040475
3	3.9394e-02	2	0.75152	0.85455	0.040017
4	2.0455e-02	3	0.71212	0.76970	0.039125
5	9.0909e-03	7	0.63030	0.71212	0.038363
6	6.0606e-03	8	0.62121	0.66061	0.037567
7	4.5455e-03	11	0.60303	0.65455	0.037466
8	3.0303e-03	13	0.59394	0.66061	0.037567
9	2.0202e-03	18	0.57879	0.67576	0.037813
10	1.5152e-03	21	0.57273	0.67576	0.037813
11	1.0000e-20	23	0.56970	0.68485	0.037956

Appendix Figure I CP plot and table for model DT-2



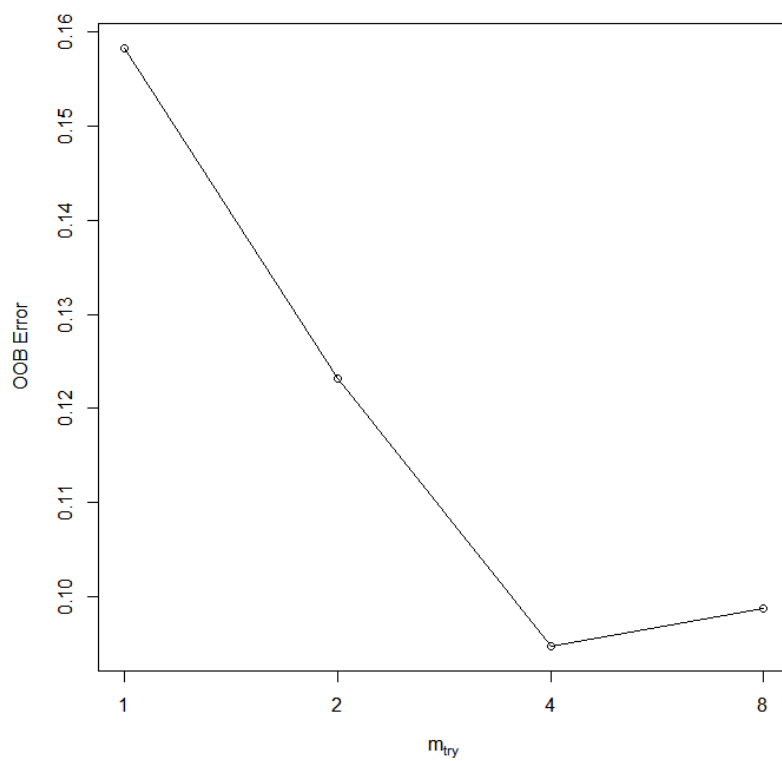
	CP	nsplit	rel error	xerror	xstd
1	1.6970e-01	0	1.00000	1.00000	0.040953
2	7.8788e-02	1	0.83030	0.90303	0.040410
3	3.9394e-02	2	0.75152	0.81818	0.039667
4	2.8788e-02	3	0.71212	0.76667	0.039088
5	1.2879e-02	5	0.65455	0.71212	0.038363
6	1.2121e-02	9	0.60303	0.70303	0.038231
7	9.0909e-03	11	0.57879	0.67879	0.037861
8	8.0808e-03	12	0.56970	0.66364	0.037617
9	6.0606e-03	15	0.54545	0.66667	0.037667
10	4.0404e-03	21	0.50909	0.65152	0.037415
11	3.0303e-03	24	0.49697	0.65758	0.037517
12	1.5152e-03	26	0.49091	0.66970	0.037716

Appendix Figure 3. CP plot and table for model DT-3

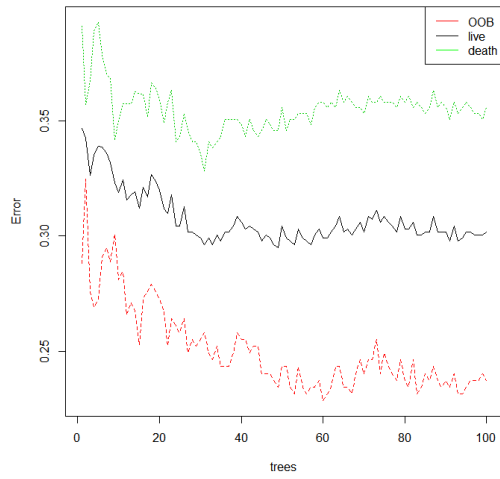
```

> tune.rf <- tuneRF(train[, -8], train$deathyes, mtryStart = 1)
mtry = 1 OOB error = 15.83%
Searching left ...
Searching right ...
mtry = 2 OOB error = 12.31%
0.2222222 0.05
mtry = 4 OOB error = 9.47%
0.2307692 0.05
mtry = 8 OOB error = 9.88%
-0.04285714 0.05

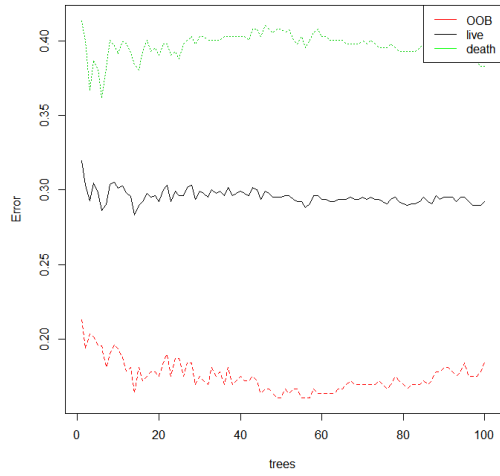
```



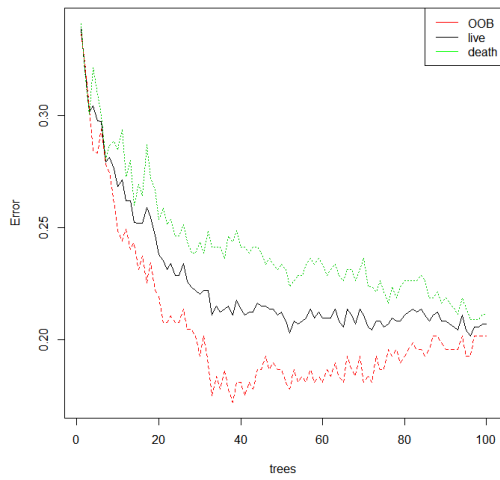
Appendix Figure 4. OOB error versus mtry (RF)



Appendix Figure 5. OOB errors according to number of trees (RF-1)



Appendix Figure 6. OOB errors according to number of trees (RF-2)



Appendix Figure 7. OOB errors according to number of trees (RF-3)

Appendix Table 24. Important variables selection of five models based on feature set 1

Factors	Variables	Model Cox-1	Model DT-1	Model RF-1	Model Bag-1	Model AdaBoost-1
Health Condition	Cancer stage	O	O	O	O	O
	Regional lymph node metastasis	O	O	O	O	O
Environmental Factors	Time since diagnosis		O	O	O	O
	Household income	O	O	O	O	O
Personal Factors	Age	O	O	O	O	O
	Sex or Gender	O	O	O	O	O

Appendix Table 25. Important variables of five models based on feature set 2

Factors	Variables	Model Cox-2	Model DT-2	Model RF-2	Model Bag-2	Model AdaBoost-2
Body function and structures	BMI(kg/m2) before operation	○	○	○	○	○
	Anxiety		○	○	○	○
	Depression		○	○	○	○
	Physical functioning					
	Role functioning	○	○	○	○	
	Dyspnea	○	○	○	○	○
	Appetite loss		○	○	○	○
	Diarrhea					
	Lung cancer specific cough					○
	Pain in chest		○	○	○	
	New possibility				○	
	Personal strength	○	○	○	○	
	Appreciation of life	○	○	○	○	
Activities	Physical activity (MET)		○	○	○	

Appendix Table 26. Important variables of five models based on feature set 3

Factors	Variables	Model Cox-3	Model DT-3	Model RF-3	Model Bag-3	Model AdaBoost-3
Health Condition	Cancer stage	O		O	O	
	Regional lymph node metastasis			O	O	
Environmental Factors	Time since diagnosis	O	O		O	
	Low household income			O	O	
Personal Factors	Age				O	O
	Sex	O	O	O	O	
	BMI(kg/m2)	O		O	O	O
	before operation			O	O	O
Body function and structures	Anxiety			O	O	O
	Depression				O	O
	Physical functioning					
	Role functioning	O			O	
	Dyspnea	O	O	O	O	O
	Appetite loss					O
	Diarrhea					
	Lung cancer specific cough					
	Pain in chest					
	New possibility				O	
Activities	Personal strength	O	O	O	O	
	Appreciation of life	O				
	Physical activity (MET)		O	O	O	

Appendix Table 27. Performance comparisons of three data mining algorithms with Cox model of the lung cancer survival prediction model with feature set 1

Feature set 1	PPV, Precision	NPV	Sensitivity (TPR)	Specificity	Accuracy (%)	AUC (95% CI)
Algorithms						
Cox-1	72.414	71.963	83.051	57.895	72.3%	0.699(0.668-0.730)
DT-1	68.235	72.789	74.359	66.460	69.1%	0.775(0.724-0.826)
RF-1	73.988	70.803	76.190	68.310	73.5%	0.821(0.775-0.867)
Bagging-1	73.529	66.667	71.839	68.531	73.5%	0.788(0.740-0.837)
AdaBoost-1	71.351	71.212	77.647	63.946	72.2%	0.819(0.774-0.864)

**Abbreviation:** DT, Decision Tree; RF, Random Forest; Ada Boost, Adjusted Boosting; PPV, Positive Predictive Value; NPV, Negative Predictive Value



Appendix Table 28. Performance comparisons of three data mining algorithms with Cox model of the lung cancer survival prediction model with feature set 2

Feature set 2	PPV, Precision	NPV	Sensitivity (TPR)	Specificity (TNR)	Accuracy (%)	AUC (95% CI)
Algorithms						
Cox-2	70.297	67.257	79.330	55.882	69.2%	0.767(0.739-0.795)
DT-2	77.698	65.169	63.529	78.912	67.5%	0.769(0.717-0.822)
RF-2	79.412	68.712	67.925	80.000	73.8%	0.789(0.728-0.830)
Bagging-2	76.437	71.329	76.437	71.329	74.1%	0.779(0.728-0.830)
AdaBoost-2	75.497	66.265	67.059	74.830	73.5%	0.785(0.735-0.835)

**Abbreviation:** DT, Decision Tree; RF, Random Forest; Ada Boost, Adjusted Boosting; PPV, Positive Predictive Value; NPV, Negative Predictive Value

Appendix Table 29. Performance comparison of three data mining algorithms with Cox model of the lung cancer survival prediction model with feature set 3

Feature set 3	PPV, Precision	NPV	Sensitivity (TPR)	Specificity (TNR)	Accuracy (%)	AUC (95% CI)
<b>Algorithms</b>						
Cox-3	76.289	75.000	83.616	65.414	76%	0.809(0.783-0.835)
DT-3	75.497	66.265	67.059	74.830	76%	0.800(0.750-0.850)
RF-3	86.164	84.962	87.261	83.704	82%	0.918(0.888-0.947)
Bagging-3	81.461	82.014	85.294	77.551	77.9%	0.834(0.789-0.878)
AdaBoost-3	84.431	79.137	82.941	80.882	85%	0.893(0.838-0.927)

**Abbreviation:** DT, Decision Tree; RF, Random Forest; Ada Boost, Adjusted Boosting; PPV, Positive Predictive Value; NPV, Negative Predictive Value

Appendix Table 30. Best fitting model from k-fold cross-validation based on Decision Tree (DT) models

Model DT-1			Model DT-2			Model DT-3		
Min criterion	Accuracy	Kappa	Min criterion	Accuracy	Kappa	Min criterion	Accuracy	Kappa
0.010	0.633	0.273	0.010	0.700	0.404	0.010	0.678	0.358
0.255	0.622	0.256	<b>0.255</b>	<b>0.709</b>	<b>0.419</b>	<b>0.255</b>	<b>0.689</b>	<b>0.379</b>
0.500	0.622	0.255	0.500	0.694	0.392	0.500	0.681	0.36
<b>0.745</b>	<b>0.642</b>	<b>0.299</b>	0.745	0.686	0.376	0.745	0.659	0.322
0.990	0.605	0.218	0.990	0.658	0.308	0.990	0.653	0.299

Appendix Table 31. Best fitting model from k-fold cross-validation based on Random Forest (RF) models

Model RF-1			Model RF-2			Model RF-3		
mtry	Accuracy	Kappa	mtry	Accuracy	Kappa	mtry	Accuracy	Kappa
2	0.651	0.305	2	0.775	0.550	2	0.783	0.564
3	0.655	0.318	<b>4</b>	<b>0.779</b>	<b>0.560</b>	<b>8</b>	<b>0.784</b>	<b>0.567</b>
<b>4</b>	<b>0.659</b>	<b>0.325</b>	7	0.768	0.538	15	0.779	0.557
5	0.646	0.296	10	0.768	0.528	21	0.776	0.551
6	0.652	0.307	13	0.750	0.502	28	0.771	0.541

Appendix Table 32. Comparisons of accuracy of hold-out sampling vs. that of k-fold cross validation

	Accuracy (%)	K-fold cross validation Accuracy (%)
Model DT-1	69.10%	64.15%
Model DT-2	67.50%	70.85%
Model DT-3	76.30%	68.94%
Model RF-1	73.50%	65.87%
Model RF-2	73.80%	77.93%
Model RF-3	82.30%	78.42%
Model Bagging-1	73.5%	61.34%
Model Bagging-2	74.1%	71.43%
Model Bagging-3	77.9%	73.11%
Model AdaBoost-1	72.2%	64.69%
Model AdaBoost-2	73.5%	72.85%
Model AdaBoost-3	84.9%	75.35%

본 학위 논문은 한국연구재단 글로벌 박사양성사업 (2016907839) 의 지원을 받았으며, 국립암센터 기관고유연구사업 (0710410 & 1010470) 의 데이터를 활용하여 작성 되었음을 밝힙니다.